Uniqueness and On-line Algorithms in Identification of Linear Dynamic Systems

Söderström, Torsten

1973

SIGILLUM UNIVERSITATIS GOTHOR. CAROLINÆ LUND. * * *

AD UTRUMQUE

1666.

LTH

REPORT 7314
APRIL 1973

Uniqueness and On-line Algorithms in Identification of Linear Dynamic Systems

TORSTEN SÖDERSTRÖM

LTH

Division of Automatic Control · Lund Institute of Technology

To Marianne

Printed in Sweden
Saldo Tryckeri AB
Flädie 1973

INTRODUCTION.

In most of the existing theory of automatic control the design of a control law requires a mathematical model of the process. There are in principal two ways to construct mathematical models of dynamic systems. One way is to use basic physical laws and another to use process identification techniques. For many types of processes no well-known physical laws are applicable or some constants are unknown, so process identification must be used. This means that the model is constructed from measured input output data. Many identification methods are proposed. Several of them are used extensively and with good results in different applications.

The purpose of this thesis is to examine some proposed methods of identification. The thesis consists of this summary and the following reports:

I. T. Söderström: On the Convergence Properties of the Generalized Least Squares Identification Method. Report 7228, Division of Automatic Control, Lund Institute of Technology, 1972. Includes an addendum. An abbreviated version is accepted as paper for the third IFAC Symposium on Identification and System Parameter Estimation, 1973.

II. K.J. Åström and T. Söderström: Uniqueness of the Maximum Likelihood Estimates of the Parameters of a Mixed Autoregressive Moving Average Process. Report 7306, Division of Automatic Control, Lund Institute of Technology, 1973.

III. T. Söderström: On the Uniqueness of Maximum Likelihood Identification for Different Structures. Report 7307, Division of Automatic Control, Lund Institute of Technology, 1973.

IV. T. Söderström: An On-Line Algorithm for Approximate Maximum Likelihood Identification of Linear Dynamic Systems. Report 7308, Division of Automatic Control, Lund Institute of Technology, 1973.

The analysis is limited to systems with a single input and a single output. The methods considered here can all be formulated as optimization problems. Since many identification schemes lead to strongly nonlinear loss functions the optimizations must be done computationally using some search routine. Since this may give only a local instead of a global optimum, the resulting model from the identification may depend on the start values of the optimization. Thus uniqueness of the methods is closely related to the number of optimum points. Investigation of the uniqueness of some different identification methods is the main topic discussed in this thesis. It is treated in parts I, II and III.

In many applications it is attractive to have an efficient on-line identification method. For adaptive control it is necessary to use an on-line algorithm for identification. In other cases it may be desirable to proceed the identification until a specified accuracy is obtained. A way to convert the maximum likelihood method into an on-line method is discussed in part IV.

Some of the results are illustrated by numerical examples. Plant measurements as well as input output data from simulated systems are used.

SOME IDENTIFICATION SCHEMES.

In order to analyse the uniqueness of a method it will be assumed that the output of the system is generated by a linear, time-invariant difference equation of the following form:

$$y(t) = \frac{B(q^{-1})}{A(q^{-1})} u(t) + \frac{C(q^{-1})}{D(q^{-1})} e(t) \qquad (1)$$

where y(t) denotes the output at time t, u(t) the input and {e(t)} white noise (a sequence of independent, equally distributed random variables). The backward shift operator is denoted by q^{-1} and the polynomial operators of (1) are of the following form:

$$\begin{aligned} A(q^{-1}) &= 1 + a_1 q^{-1} + \dots + a_n q^{-n} \\ B(q^{-1}) &= \phantom{1 + {}} b_1 q^{-1} + \dots + b_n q^{-n} \\ C(q^{-1}) &= 1 + c_1 q^{-1} + \dots + c_n q^{-n} \\ D(q^{-1}) &= 1 + d_1 q^{-1} + \dots + d_n q^{-n} \end{aligned} \qquad (2)$$

Special cases of (1) will often be considered. Some of the polynomials may be substituted by 1 or there may be some constraints on the polynomial coefficients. This will reduce the number of parameters describing the system. After this possible reduction the parameters are collected in a vector θ.

The class of systems considered can without serious difficulties be extended to include systems with multiple inputs and one output, systems with a delay in the transfer function and systems with polynomials of different degrees.

The purpose of identification is to estimate the parameter vector θ using measurements y(1), ..., y(N), u(1), ..., u(N). When the noise e(t) is gaussian the methods treated in the thesis can be interpreted as the maximum likelihood method applied to the particular type of system. Computationally the identification methods minimize the loss function $V_N(\hat{\theta})$ defined by

$$V_N(\hat{\theta}) = \frac{1}{N} \sum_{t=1}^{N} \varepsilon^2(t;\hat{\theta}) \qquad (3)$$

$$y(t) = \frac{\hat{B}(q^{-1})}{\hat{A}(q^{-1})} u(t) + \frac{\hat{C}(q^{-1})}{\hat{D}(q^{-1})} \varepsilon(t;\hat{\theta}) \qquad (4)$$

The polynomials $\hat{A}(q^{-1})$, ..., $\hat{D}(q^{-1})$ denote estimates of $A(q^{-1})$, ..., $D(q^{-1})$ and the coefficients are collected in a vector $\hat{\theta}$. Equation (4) will be called the model of the system.

The methods have an interpretation also if e(t) is not gaussian. The residuals $\varepsilon(t;\hat{\theta})$ are the one step prediction errors when the model (4) is used. Thus the methods mean that an estimated variance of the one step prediction error is minimized.

It is convenient to require that the polynomials describing the system are such that $A(z)$, $C(z)$ and $D(z)$ have all zeros outside a circle $|z| = r > 1$, where r may be close to 1. Only points $\hat{\theta}$ which fulfil the corresponding condition for $\hat{A}(z)$, $\hat{C}(z)$ and $\hat{D}(z)$ are considered in the minimization of $V_N(\hat{\theta})$. The physical interpretation of these conditions is that the deterministic part of the system is asymptotically stable and that the disturbances and the residuals have finite variances.

Strictly speaking the loss function must include the initial values of (4) and it is to be minimized with respect to these as well. However, in the analysis of $V_N(\hat{\theta})$ asymptotic theory is used and when N is large the initial values have small influence on the loss function according to the assumptions on the polynomials. This is the reason why the effect of the initial values is generally neglected in the whole thesis.

The following different cases of (4) are considered in the thesis. The restrictions on the polynomials are given for the models, but in the analysis of uniqueness it is generally assumed that the corresponding restrictions are valid for the system.

1. The well-known least squares (LS) method which corresponds to

$$\hat{A}(q^{-1}) = \hat{D}(q^{-1}), \quad \hat{C}(q^{-1}) = 1$$

2. The LS method has been extended by Clarke (1967) to the generalized least squares (GLS) method. Different versions of this method exist. For version 1, here called GLS1, see I ,

$$\hat{C}(q^{-1}) = 1, \quad \hat{D}(q^{-1}) = \hat{A}(q^{-1})\hat{F}(q^{-1})$$

which implies that the degree of $\hat{D}(q^{-1})$ is greater than n.

3. The second version, GLS2, does not immediately correspond to the above scheme. It corresponds to a model of the form, see I

$$\hat{C}(q^{-1}) = 1, \quad \hat{D}(q^{-1}) = \hat{A}(q^{-1}) \prod_{i=1}^{\infty} \hat{F}_i(q^{-1})$$

4. The third version, GLS3, see Clarke (1973), is given by

$$\hat{C}(q^{-1}) = 1$$

5. The time series case implies that there is no input, i.e.

$$\hat{A}(q^{-1}) = 1, \quad \hat{B}(q^{-1}) = 0$$

6. If the noise is white measurement noise

$$\hat{C}(q^{-1}) = 1, \quad \hat{D}(q^{-1}) = 1$$

This case will be called ML1.

7. The model used by Åström-Bohlin (1965) requires

$$\hat{A}(q^{-1}) = \hat{D}(q^{-1})$$

It will be denoted by ML2.

8. Bohlin (1970) and Box-Jenkins (1970) treat a model of the type (4) with no specific restrictions. It will be called ML3.

The model GLS3 is a simple special case of ML3 and the analysis of the latter will immediately be applicable for GLS3. These two models, however, are computed using quite different numerical methods. They are therefore treated as two separated cases.

UNIQUENESS.

In the parts I, II and III the uniqueness of the methods are discussed. The number of local minimum points of the loss function is examined. It is well-known that using real data the loss function $V_N(\hat{\theta})$ may have more than one local minimum. In the analysis it will be assumed that the data are actually generated by an equation of a correct structure. The function $V_N(\hat{\theta})$ is a stochastic variable and in order not to involve probability theory in this examination asymptotic theory is used. It is shown in I that $V_N(\hat{\theta})$ under suitable conditions has a limit $W(\hat{\theta})$ with probability one when the number of samples tends to infinity. If $\varepsilon(t;\hat{\theta})$ does not contain any deterministic part then $W(\hat{\theta}) = E\varepsilon^2(t;\hat{\theta})$.

Some general properties of the uniqueness of the different methods (with exclusion of GLS2) are stated. The loss function $W(\hat{\theta})$ has always a global minimum point in $\hat{\theta} = \theta$. Further there is no unique global optimum if the degrees of the polynomials in (4) are chosen too high. If the degrees are chosen in a proper way there will be a unique global minimum if the input is persistently exciting of sufficiently high an order. In the following it is assumed that the degrees of the polynomials are chosen properly.

As an attempt to analyse the uniqueness, the stationary points of $W(\hat{\theta})$ are considered, that is the solutions of

$$W'(\hat{\theta}) = 0 \tag{5}$$

In the LS case this equation is linear and the analysis is trivial. In other cases the equation is non-linear and in general very hard to solve in a straightforward way. There are two technical difficulties in the analysis of (5). The first is that $W(\hat{\theta})$ in general consists of two parts, one due to the input and the other due to noise. The analysis would be somewhat simpler if only one part was considered. The second difficulty is that it is highly desirable to treat (5) for a general class of input signals. The second difficulty may be avoided by choosing u(t) as white noise. For the ML1 case this attempt leads to reasonable calculations.

In the analysis of (5) two different approaches have been used.

The first approach is considered in II and III. The equation (5) is rewritten in the form

$$P(\hat{\theta})(\hat{\theta}-\theta) = 0$$

and for special cases this transformation has been done so that $P(\hat{\theta})$ is non-singular for all $\hat{\theta}$. This implies that $\hat{\theta} = \theta$ is the only solution of (5).

In I the second approach is analysed. The equation (5) is treated for very high and very small values of the signal-to-noise ratio.

It turns out that for some cases it is suitable to write (5) in the form

$$W_1'(x,z) + \delta W_2'(x,z) = 0$$

where $\hat{\theta}$ is decomposed as (x,z) and δ is a small number. For the specific new form of the equation it turns out that the solutions of $W_1'(x,z) = 0$ are $x = 0$, z arbitrary. One could expect that the solutions of (5) for small values of δ are approximately given by $x = 0$ and z a solution of $W_2'(0,z) = 0$. This idea is examined and conditions are given for which this is true. Clearly $W_1''(0,z)$ is singular and thus the ordinary inverse function theorem cannot be applied.

In the time series case none of the described difficulties occur. It is possible to obtain an analytic representation for the asymptotic loss function, namely

$$W(\hat{\theta}) = \frac{1}{2\pi i}\oint \frac{\hat{A}(z)C(z)}{A(z)\hat{C}(z)}\,\frac{\hat{A}(z^{-1})C(z^{-1})}{A(z^{-1})\hat{C}(z^{-1})}\,\frac{dz}{z}$$

where the integration path is the unit circle. The equation (5) is evaluated with residue calculus and in principle straightforward calculations lead to the result. For this case it is shown that $W(\hat{\theta})$ has a unique local minimum point at $\hat{\theta} = \theta$.

It is shown in III that the first approach is useful to the analysis of the case ML1. It has been proven that if the degree of $A(q^{-1})$ is one ($B(q^{-1})$ may have an arbitrary degree) then $\hat{\theta} = \theta$ is the only stationary point of $W(\hat{\theta})$. In the special case when the input is white noise, uniqueness is proven for an arbitrary degree of $A(q^{-1})$.

In III it is also shown how the results for ML1 and for the time series case can be applied to the ML3 case. For this model $\hat{\theta} = \theta$ is the only stationary point if the degree of $A(q^{-1})$ is one. This result holds for GLS3 as well.

The second approach is applied to GLS1 in I and to ML2 in III. It turns out that under suitable conditions the method ML2 gives unique models. The essential property which is required to prove uniqueness is that the signal to noise ratio is either sufficiently small or sufficiently large. For the method GLS1, however, the number of local minimum points of $W(\hat{\theta})$ depends on the signal to noise ratio. When this number is sufficiently large $\hat{\theta} = \theta$ is the only local minimum point. For small values of the ratio there are in general several minimum points. The nonuniqueness of the GLS1 method is illustrated in I using different plant measurements and simulated input output data. The loss functions have in these cases at least two minimum points although the signal to noise ratio is mostly reasonable.

The method called GLS2 is treated in I. It is shown by counter-examples that this method can give wrong values of the estimates if the signal to noise ratio is large.

ON-LINE VERSIONS OF IDENTIFICATION METHODS.

It is sometimes desirable to have an on-line algorithm for identification of a system. It is easy to convert the LS method into a recursive method, see e.g. Åström (1968). An on-line version of the GLS method is described in Hasting-James and Sage (1969). They use another approach than the one given here. In IV a more general way of generating on-line versions of identification methods is given. For the LS case it reduces to the usual recursive LS algorithm.

The main idea is now described. Let $\hat{\theta}_N$ denote the minimum point of $V_N(\hat{\theta})$. The estimate $\hat{\theta}_{N+1}$ is found by minimization of $V_{N+1}(\hat{\theta})$. If it is assumed that $\hat{\theta}_{N+1}$ is so close to $\hat{\theta}_N$ that one Newton-Raphson iteration is sufficient, then

$$\hat{\theta}_{N+1} = \hat{\theta}_N - V''_{N+1}(\hat{\theta}_N)^{-1}V'_{N+1}(\hat{\theta}_N)^T \qquad (6)$$

Using the relation $V_{N+1}(\hat{\theta}) = V_N(\hat{\theta}) + \varepsilon^2(N+1;\hat{\theta})$ the derivatives in (6) can be expressed in derivatives of $V_N(\hat{\theta})$ and $\varepsilon(N,\hat{\theta})$ evaluated at $\hat{\theta}_N$. Under certain assumptions, which all are fulfilled exactly for the LS case, the algorithm gets the following form:

$$\hat{\theta}_{N+1} = \hat{\theta}_N - P_{N+1}\varepsilon'(N+1;\hat{\theta}_N)^T\varepsilon(N+1;\hat{\theta}_N) \qquad (7)$$

$$P_{N+1} = P_N - P_N\varepsilon'(N+1;\hat{\theta}_N)^T\varepsilon'(N+1;\hat{\theta}_N)P_N/\{1+\varepsilon'(N+1;\hat{\theta}_N)P_N\varepsilon'(N+1;\hat{\theta}_N)^T\} \qquad (8)$$

Special interest has been paid to the ML2 model. Equations are derived which compute $\varepsilon'(N+1;\hat{\theta}_N)$ and $\varepsilon(N+1;\hat{\theta}_N)$ recursively in an approximate way. With some additional simplifications the algorithm reduces to the method proposed by Young (1970).

The algorithm applied in a straightforward way does not work satisfactorily on simulated data. In IV it is demonstrated that the algorithm can be improved considerably by various tricks like restarts and by processing of a number of measurements without change of $\hat{\theta}$. Simulated data as well as plant measurements are used for illustration of the algorithm.

ACKNOWLEDGEMENTS.

I am very glad to express my gratitude to my supervisor Professor Karl Johan Åström. For a great deal of my work I used his preliminary analysis, and his guidance throughout this work has been of invaluable help.

I am also indebted to several persons at the Division of Automatic Control. In particular I want to sincerely thank Tekn.lic. Ivar Gustavsson for many stimulating discussions on all parts of the thesis.

It is also a great pleasure to thank the people who have helped me with practical things in the preparation of the thesis. Mrs. Gudrun Christensen and Miss Kerstin Palmqvist have done an excellent job at the type-writers. Mr. Bengt Lander and Mrs. Birgitta Tell have successfully prepared the figures.

The plant data used in this thesis belongs to a data bank at the Division of Automatic Control. The particular data-sets have been supplied by AB Atomenergi and the OECD Halden Reactor Project (nuclear reactor data), the National Physical Laboratory, Teddington, England (distillation column data) and Tekn.lic. Bo Leden (heat rod data). The permission to use these data are gratefully acknowledged.

The publication of the reports has been partially supported by the Swedish Board of Technical Development under contract 72-202/U 137.

REFERENCES.

Åström, K.J. (1968).
Lectures on the Identification Problem - The Least Squares Method. Report 6806, Division of Automatic Control, Lund Institute of Technology.

Åström, K.J. - Bohlin, T. (1965).
Numerical Identification of Linear Dynamic Systems from Normal Operating Records. Paper, IFAC Symposium on Theory of Self-Adaptive Systems, Teddington, England. In Theory of Self-Adaptive Control Systems (Ed. P.H.Hammond), Plenum Press, New York.

Bohlin, T. (1970).
On the Maximum Likelihood Method of Identification. IBM J. Res. and Dev., 14, No. 1, p. 41 - 51.

Box, G.E.P. - Jenkins, G.M. (1970).
Time Series Analysis Forecasting and Control. Holden-Day, San Francisco.

Clarke, D.W. (1967).
Generalized Least Squares Estimation of the Parameters of a Dynamic Model. 1st IFAC Symposium on Identification in Automatic Control Systems, Prague.

Clarke, D.W. (1973).
Experimental Comparison of Identification Methods. Paper submitted to UKAC Conference, Bath.

Hasting-James, R. - Sage, M.W. (1969).
Recursive Generalized-Least-Squares Procedure for On-line Identification of Process Parameters. Proc. IEE, Vol. 116, No. 12, P. 2057-2062.

Young, P.C. (1970).
An Extension of the Instrumental Variable Method for Identification of a Noisy Dynamic Process. Univ. of Cambridge, Dep. of Eng., Technical note CN/70/1.

ON THE CONVERGENCE PROPERTIES OF THE GENERALIZED LEAST SQUARES IDENTIFICATION METHOD

T. Söderström

ABSTRACT.

Modelling of a discrete time system is often made by parametric identification. A linear difference equation is adapted to the dynamics of the system. The parameters of the equation can easily be estimated by the least squares method. This method has several advantages, but if the residuals are correlated, the estimates are biased. The method of generalized least squares proposed by Clarke is constructed to overcome this difficulty. This method is an iterative procedure. The dynamics of the system and the correlation of the residuals are estimated alternately.

The purpose of this report is to present an analysis of the convergence properties of the generalized least squares method. Two different variants are examined. They correspond to different ways of estimating the correlation of the residuals. It is shown that one of those variants is equivalent to a maximization of the likelihood function of the problem, when suitable assumptions are made. In this case the possible result of the method is closely related to the number of local minimum points of a corresponding loss function. Under the assumption of suitable regularity conditions of the input signal and the system dynamics the following is theoretically shown in the report.

For every given system the minimization gives the true values of the parameters if the signal to noise ratio is high enough. It is further shown that the minimization may give wrong values of the parameters if the signal to noise ratio is low enough. In this case the loss function has no unique local minimum point.

The second variant is the one proposed by Clarke. By counterexamples it is shown that also this variant may give wrong estimates for high noise levels.

The existence of wrong parameter estimates is illustrated by numerical examples. Plant measurements as well as simulated systems are used.

TABLE OF CONTENTS

I. INTRODUCTION

1.1 The structure of the system.

Consider a dynamic process. A sequence of inputs $\{u(t)\}$ and corresponding outputs $\{y(t)\}$ are given from an experiment. The purpose of an identification is to fit a mathematical model to the given data. This can be done in many ways. A good survey of different identification methods is given in [4].

In order to develop some theory it is assumed that the process is governed by some equation. The process given by this equation will be called the system in this report, while the model refers to the equation obtained in some way from the given data.

Assume that the system is linear, discrete, time invariant and of finite order. If the disturbances can be represented by stationary random processes, the system can in general be represented by

$$A(q^{-1})y(t) = B(q^{-1})u(t) + v(t) \tag{1.1}$$

where $y(t)$ is the output at time t, $u(t)$ the input at time t and $v(t)$ a stationary stochastic process. q^{-1} is the backward shift operator and

$$A(q^{-1}) = 1 + a_1q^{-1} + \dots + a_nq^{-n} \tag{1.2}$$

$$B(q^{-1}) = b_1q^{-1} + \dots + b_nq^{-n} \tag{1.3}$$

It is assumed that the system is asymptotically stable.

For simplicity introduce the following conventions

i) $e(t)$ is always denoting white noise (a sequence of independent, equally distributed random variables with zero mean

ii) σ^2 denotes the variance of $Ee^2(t)$

iii) S denotes the ratio $\frac{Eu^2(t)}{\sigma^2}$, which is proportional to the signal to noise ratio.

In the following it will be assumed that the noise v(t) can be expressed as

$$v(t) = H(q^{-1})e(t) \tag{1.4}$$

where $H(q^{-1})$ is a stable filter and e(t) white noise.

Introduce the matrix notations

$$Y = \begin{bmatrix} y(n+1) \\ \cdot \\ \cdot \\ \cdot \\ y(N+n) \end{bmatrix} \qquad V = \begin{bmatrix} v(n+1) \\ \cdot \\ \cdot \\ \cdot \\ v(N+n) \end{bmatrix}$$

$$\phi = \begin{bmatrix} -y(n).. & -y(1) & u(n)... & u(1) \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ -y(N+n-1)... & -y(N) & u(N+n-1)... & u(N) \end{bmatrix}$$

$$\theta = \begin{bmatrix} a_1 \\ \cdot \\ \cdot \\ a_n \\ b_1 \\ \cdot \\ \cdot \\ b_n \end{bmatrix}$$

(1.1) can be written as

$$Y = \phi\theta + V \tag{1.5}$$

where N is arbitrary.

1.2 The least squares method

The least squares (LS) estimate $\hat{\theta}_{LS}$ of θ is obtained by minimizing

$$V_{LS}(\hat{\theta}) = ||Y - \phi\hat{\theta}||^2 = (Y - \phi\hat{\theta})^T(Y - \phi\hat{\theta})$$

with the well-known solution

$$\hat{\theta}_{LS} = \theta + (\phi^T\phi)^{-1}\phi^T V \qquad (1.6)$$

assuming that the inverse exists.

Åström has shown [1] that this method gives consistent estimates if v(t) is white noise.

Correlated noise causes biased estimates. The generalized least squares (GLS) method introduced by Clarke [8] is intended to overcome this situation.

1.3 The Markov estimate

Introduce the symmetric matrix R, which is assumed to be non-singular

$$R = \begin{bmatrix} r_v(0)\ldots & r_v(N+1) \\ \ddots & \\ & r_v(0) \end{bmatrix}$$

$r_v(\tau)$ denotes the covariance function of the noise v(t).

If R is known the Markov estimate $\hat{\theta}_M$ of θ is obtained by minimizing

$$V_M(\hat{\theta}) = ||Y-\phi\hat{\theta}||^2_{R^{-1}} = (Y-\phi\theta)^T R^{-1}(Y-\phi\theta)$$

with the result

$$\hat{\theta}_M = \theta + (\phi^T R^{-1} \phi)^{-1} \phi^T R^{-1} v \tag{1.7}$$

which is a consistent estimate.

This follows from the consistency of the LS estimate as shown below.

It is believed that the following description of the Markov estimation besides proving consistence will give some more insight in the method and motivation for the generalized least squares method (introduced in the next section) as well.

From the relation (1.4)

$$v = He \tag{1.8}$$

where

$$H = \begin{bmatrix} 1 & & \\ h_1 & & \\ \vdots & 0 & \\ h_{N-1} & h_1 & 1 \end{bmatrix}$$

and

$$e = \begin{bmatrix} e(n+1) \\ \vdots \\ e(N+n) \end{bmatrix}$$

Define the filter $F(q^{-1})$ by

$$F(q^{-1}) = H(q^{-1})^{-1} \tag{1.9}$$

and form a corresponding matrix

$$F = \begin{bmatrix} 1 & & & \\ f_1 & & & \\ \vdots & & 0 & \\ \vdots & & & \\ f_{N-1} & & f_1 & 1 \end{bmatrix} \tag{1.10}$$

(1.9) can then be written

$$F = H^{-1} \tag{1.11}$$

From (1.8) it follows that

$$R = Evv^T = \sigma^2 HH^T$$

and invoking (1.11)

$$R^{-1} = \frac{1}{\sigma^2} F^T F \tag{1.12}$$

Introduce the filtered signals

$$\begin{aligned} y^F(t) &= F(q^{-1}y(t) \\ u^F(t) &= F(q^{-1})u(t) \end{aligned} \tag{1.13}$$

or in matrix language

$$\phi^F = F\phi, \; Y^F = FY \tag{1.14}$$

Then $V_M(\hat{\theta}) = \frac{1}{\sigma^2}(Y^F - \phi^F\hat{\theta})^T(Y^F - \phi^F\hat{\theta})$ (1.15)

From (1.5), (1.8), (1.11) and (1.14)

$$Y^F = \phi^F\theta + e \tag{1.16}$$

From (1.15) and (1.16) it is seen that the consistency of $\hat{\theta}_M$ follows from the consistency of the LS estimate.

In figure 1 the configuration adapted to LS is shown

v(t) = e(t) (white noise)

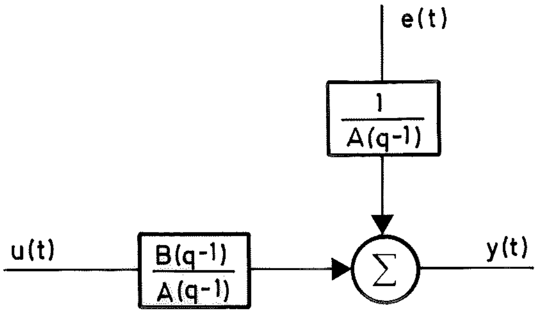


Figure 1

Figure 2 shows the general situation corresponding to (1.1)

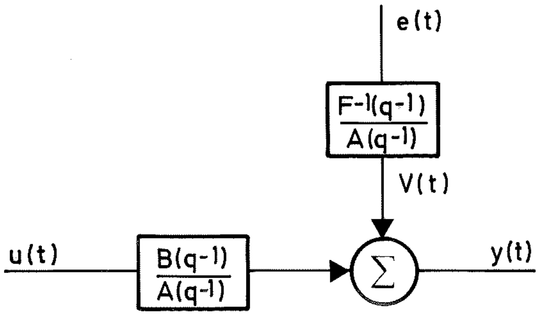

e(t)
$\frac{F^{-1}(q^{-1})}{A(q^{-1})}$
V(t)
u(t)
$\frac{B(q^{-1})}{A(q^{-1})}$
Σ
y(t)

Figure 2

This system can, however, also be represented by figure 3, where the filter $F(q^{-1})$ has been moved and the filtered signals $u^F(t)$ and $y^F(t)$ have been introduced.

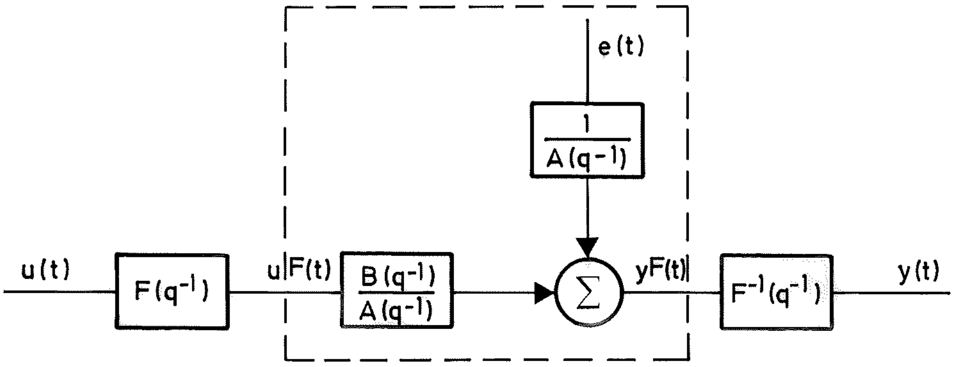


Figure 3

If R and then the filter $F(q^{-1})$ are known, $u^F(t)$ and $y^F(t)$ are easily obtained and it is sufficient to deal with the framed part of the system. This part, however, is quite similar to figure 1, thus indicating the consistency of the Markov estimate.

1.4. The generalized Least Squares method. Two versions.

The assumption of R known is highly unrealistic. In the general least squares (GLS) method θ and R are both estimated in an iterative way.

1. Guess a covariance matrix R_k.

2. Compute $\hat{\theta}_k$ from (1.7) with $R = R_k$.

3. Evaluate the residuals $\varepsilon_k = Y-\phi\hat{\theta}_k$ and use them to estimate a new covariance matrix R_{k+1}.

4. Put k=k+1 and repeat from 2 until the estimate converges.

In this report two versions of the generalized least squares method are treated. In both versions the estimates of R are obtained by fitting an autoregression to the residuals.

Version 1:

This version can be described by the following scheme.

1. Guess a filter $\hat{C}_k(q^{-1}) = 1 + \hat{c}_{k1}q^{-1}+\ldots+\hat{c}_{kn}q^{-n}$

2. Compute $y_k^F(t)$ and $u_k^F(t)$ from

$$y_k^F(t) = \hat{C}_k(q^{-1})y(t) \tag{1.17}$$

$$u_k^F(t) = \hat{C}_k(q^{-1})u(t)$$

and determine $\hat{\theta}_k$ by applying LS to the model

$$\hat{A}_k(q^{-1})y_k^F(t) = \hat{B}_k(q^{-1})u_k^F(t) + e(t)$$

3. Evaluate the residuals

$$\varepsilon_k(t) = \hat{A}_k(q^{-1})y(t) - \hat{B}_k(q^{-1})u(t) \tag{1.18}$$

Determine $\hat{C}_{k+1}(q^{-1})$ by fitting an autoregression to the residuals.

4. Put k=k+1 and repeat from 2 until convergence.

Clearly, this version corresponds to the model

$$\hat{A}(q^{-1})y(t) = \hat{B}(q^{-1})u(t) + \frac{1}{\hat{C}(q^{-1})} e(t) \tag{1.19}$$

with e(t) white noise.

Version 2:

This version coincides with Clarkes original proposal [8]. The iteration scheme is the following.

0. Put $y_o^F(t) = y(t)$, $u_o^F(t) = u(t)$, $k=1$

1. Guess a filter $\hat{C}_k(q^{-1}) = 1 + \hat{c}_{k1}q^{-1}+\ldots+ \hat{c}_{kn}q^{-n}$

2. Compute $y_k^F(t)$ and $u_k^F(t)$ from

$$\begin{aligned} y_k^F(t) &= \hat{C}_k(q^{-1})y_{k-1}^F(t) \\ u_k^F(t) &= \hat{C}_k(q^{-1})u_{k-1}^F(t) \end{aligned} \tag{1.17'}$$

and determine $\hat{\theta}_k$ by applying LS to the model

$$\hat{A}_k(q^{-1})y_k^F(t) = \hat{B}_k(q^{-1})u_k^F(t) + e(t)$$

3. Evaluate the residuals

$$\varepsilon_k(t) = \hat{A}_k(q^{-1}y_k^F(t) - \hat{B}_k(q^{-1})u_k^F(t) \qquad (1.18')$$

and determine a new filter $\hat{C}_{k+1}(q^{-1})$ by fitting an autoregression to the residuals.

4. Put k=k+1 and repeat from 2 until convergence.

With this version a successful iteration procedure ends when

$$\hat{C}_k(q^{-1}) \approx 1$$

The corresponding model is

$$\hat{A}(q^{-1})y(t) = \hat{B}(q^{-1})u(t) + \frac{1}{\prod_{k=1}^{\infty} \hat{C}_k(q^{-1})} e(t) \qquad (1.20)$$

For both the versions of GLS it is of course not necessary that the orders of the operators $\hat{A}$, $\hat{B}$ and $\hat{C}$ are the same. In this report the orders will in general be assumed to be the same, but the generalization is trivial.

The second version may be better if the noise v(t) is not generated as an autoregression. It will be shown, however, that both versions may fail (give biased estimates) at high noise levels.

The GLS method has some similarity with the repeated LS method as pointed out in [4].

In the repeated LS method (LS with successively higher order of the model) it is hoped that the A and B polynomials will have some factors in common. These factors are due to the correlation of the present noise.

In the GLS method there are always factors in common. To realized that, (1.19) is rewritten in the form

$$[\hat{A}(q^{-1})\hat{C}(q^{-1})]y(t) = [\hat{B}(q^{-1})\hat{C}(q^{-1})]u(t) + e(t)$$

The GLS method can thus be interpreted as a LS method with the constraint that the A and B polynomials have common factors.

In order to closer examine the properties of the two versions, the nature of the noise v(t) or the covariance function $r_v(\tau)$ must be specified.

Some results in this report require only

$$v(t) = H(q^{-1})e(t)$$

where $H(q^{-1})$ is a stable filter and e(t) is white noise.

Sometimes special interest will be paid to the following filter of finite order

$$H(q^{-1}) = \frac{1}{C(q^{-1})}$$

where

$$C(q^{-1}) = 1 + c_1q^{-1}+\ldots+c_nq^{-n}$$

has all zeros outside the unit circle.

In these cases obviously

$$v(t) = \frac{1}{C(q^{-1})} e(t) \tag{1.21}$$

The reason for a study of (1.21) is its similarity in structure with the model (1.19).

It will be shown that under suitable regularity conditions on the input signal and the system dynamics the first version of the GLS method will always give consistent estimates, if the signal to noise ratio is high enough. However, if the noise level is high enough this version can give asymptotically biased estimates. It will also be shown that the second version can give biased estimates if the signal to noise ratio is low. All results hold asymptotically when the number of data tends to infinity.

II. MATHEMATICAL PRELIMINARIES

2.1. Ergodic properties of time series

It is the purpose to develop results which are valid as the number of data tends to infinity.

The least squares estimate $\hat{\theta}_{LS}$ (1.4) can be written

$$\hat{\theta}_{LS} = \theta + \left(\frac{\phi^T\phi}{N}\right)^{-1}\left(\frac{\phi^T v}{N}\right)$$

The elements of the matrices $\frac{\phi^T\phi}{N}$ and $\frac{\phi^T v}{N}$ are sample covariances. It is valuable to know when these sample covariances converge as $N\to\infty$, and in case of convergence the limits too.

The questions are answered by ergodic theory. Some results of this nature are collected in Appendix A.

The main result is the following.

Theorem 2.1: Consider the system

$$y(t) = G(q^{-1})u(t) + H(q^{-1})\ e(t)$$

where

$G(q^{-1})$ and $H(q^{-1})$ are asymptotically stable filters of finite orders.

$e(t)$ is white noise with finite fourth moment and independent of $u(t)$

$$u(t) = u_1(t) + u_2(t)$$

$u_1(t)$ deterministic and almost periodic, that is to every $\varepsilon > 0$ there is a periodic function $u_1'(t)$ such that

$$|u_1(t) - u_1'(t)| < \varepsilon \quad \text{all } t$$

$u_2(t) = F(q^{-1})\ v(t)$ with $F(q^{-1})$ an asymptotically stable filter of finite order and v(t) white noise with finite fourth moment.

Let further $D_1(q^{-1})$ and $D_2(q^{-1})$ be asymptotically stable filters of finite orders.

Then

$$\lim \frac{1}{n} \sum_{t=1}^{n} (D_1(q^{-1})y(t) + D_2(q^{-1})u(t)) \begin{bmatrix} y(t) \\ u(t) \end{bmatrix}$$

$$= E(D_1(q^{-1})y(t) + D_2(q^{-1})u(t))\begin{bmatrix} y(t) \\ u(t) \end{bmatrix} \tag{2.1}$$

with probability one and in mean square.

If x(t) is deterministic, E x(t) denotes $\lim\limits_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} x(t)$.

2.2. Persistently Exciting Signals.

Definition 2.1: u(t) is said to be persistently exciting of order n if

i) $\lim\limits_{N\to\infty} \frac{1}{N} \sum\limits_{t=1}^{N} u(t) = \bar{u}$ and $\lim\limits_{N\to\infty} \frac{1}{N} \sum\limits_{t=1}^{N} [u(t)-\bar{u}][u(t+\tau)-\bar{u}] = r_u(\tau)$

exist and

ii) the n by n symmetric matrix

$$R_u = \begin{bmatrix} r_u(0) & r_u(1) \ldots\ldots & r_u(n-1) \\ & & \\ & & r_u(1) \\ & & r_u(0) \end{bmatrix}$$

is positive definite.

Some simple properties of persistently exciting signals and a characterization of this concept in the frequency domain is given in [15]. In this report the following properties will be used (proved in [15]).

Lemma 2.1: u(t) is persistently exciting of order n if and only if the spectral density corresponding to the sample covariance function is non zero (in distributive sense) in at least n different points.

If u(t) is periodic, the spectral density will be discrete and consist of a number of δ-functions. The distribution $\delta(x)$ is here considered as non zero in x = 0.

Corr: Let $y(t) = H(q^{-1})u(t)$. If u(t) is persistently exciting of order n and $H(q^{-1})$ is stable and has no zeros on the unit circle, then y(t) is persistently exciting of order n.

A simple application of the definition is made in

Lemma 2.2: Let $y(t) = H(q^{-1})u(t)$ $\quad H(q^{-1}) = \sum_{i=0}^{n-1} h_i q^{-i}$

i) If $y(t) \equiv 0$ with probability one and $u(t)$ is persistently exciting, then $h_i = 0$ $i = 0,\ldots,n-1$

ii) If $u(t)$ is not persistently exciting of order n, then there exists $H(q^{-1}) \not\equiv 0$ such that $y(t) \equiv 0$ with probability one.

Proof:

$$Ey^2(t) = \begin{bmatrix} h_o \ldots h_{n-1} \end{bmatrix} \begin{bmatrix} r_u(0)\ldots\ldots & & r_u(n-1) \\ & & \\ & & r_u(0) \end{bmatrix} \begin{bmatrix} h_o \\ \\ h_{n-1} \end{bmatrix}$$

$y(t) = 0$ with probability one if and only if $Ey^2(t) = 0$.

i) $Ey(t)^2 = 0$ and R_u non singular implies $h_i = 0$ $i = 0,\ldots n-1$

ii) R_u is singular. Take the vector

$$\begin{bmatrix} h_o \\ \vdots \\ h_{n-1} \end{bmatrix}$$

in the null space of R_u. Then $E\, y(t)^2 = 0$.

Q.E.D.

2.3. The system covariance matrix

Consider the undisturbed linear system

$$y(t) = K(q^{-1})u(t)$$

Definition 2.2: The system covariance matrix of order 2k is understood as the 2k by 2k symmetric matrix

$$R = \begin{bmatrix} R_y & R_{yu} \\ R_{uy} & R_u \end{bmatrix}$$

$$= \begin{bmatrix} r_y(0)\ldots & r_y(k-1) & r_{yu}(0)\ldots & r_{yu}(k-1) \\ & r_y(0) & r_{yu}(1-k)\ldots & r_{yu}(0) \\ & & r_u(0) & r_u(k-1) \\ & & & \\ & & & r_u(0) \end{bmatrix}$$

$$= \lim_{N\to\infty} \frac{1}{N} \sum_{t=n+1}^{n+N} \begin{bmatrix} y(t-1) \\ \\ y(t-k) \\ u(t-1) \\ u(t-k) \end{bmatrix} [y(t-1)\ldots\ y(t-k)u(t-1)\ldots\ u(t-k)]$$

Lemma 2.1. Let R be the system covariance matrix of order k of

$$y(t) = K(q^{-1})u(t)$$

Then

$$x^TRx = r_\varepsilon(0)$$

with

$$\varepsilon(t) = F(q^{-1})y(t) + G(q^{-1})u(t)$$

$$F(q^{-1}) = \sum_{1}^{k} f_i q^{-i}, \quad G(q^{-1}) = \sum_{1}^{k} g_i q^{-i}$$

$$x = [f_1 \ldots f_k g_1 \quad g_k]^T$$

Proof: Straight forward calculations give

$$x^T R x = \lim_{N\to\infty} \frac{1}{N} x^T \sum_{t=n+1}^{n+N} \begin{bmatrix} y(t-1) \\ \vdots \\ u(t-k) \end{bmatrix} [y(t-1) \ldots u(t-k)]x$$

$$= \lim_{N\to\infty} \frac{1}{N} \sum_{t=n+1}^{n+N} ([f_1 \ldots . f_k g_1 \ldots g_k] \begin{bmatrix} y(t-1) \\ y(t-k) \\ u(t-1) \\ u(t-k) \end{bmatrix})^2$$

$$= \lim_{N\to\infty} \frac{1}{N} \sum_{t=n+1}^{n+N} \varepsilon^2(t) = r_\varepsilon(0)$$

Q.E.D.

Theorem 2.2.: Let the controllable, asymptotically stable system

$$A(q^{-1})y(t) = B(q^{-1})u(t)$$

be of order n.

Consider the system covariance matrix R of order 2k.

i) Assume that $k \leq n$. If u(t) is persistently exciting of order n+k, then R is positive definite.

ii) Assume that $k > n$. If u(t) is persistently exciting of order n+k, then R is singular (positive semidefinite). Further the null space of R is spanned by vectors of the form

$$x = \begin{bmatrix} f_1 \\ \\ f_k \\ g_1 \\ \\ g_k \end{bmatrix} \tag{2.2}$$

where f_i and g_i fulfil the relations

$$F(q^{-1}) = \sum_{i=1}^{k} f_i q^{-i} = A(q^{-1})\ L(q^{-1}) \tag{2.3a}$$

$$G(q^{-1}) = \sum_{i=1}^{k} g_i q^{-i} = -B(q^{-1})\ L(q^{-1}) \tag{2.3b}$$

$$L(q^{-1}) = \sum_{i=1}^{k-n} l_i q^{-i} \quad \text{is arbitrary} \tag{2.4}$$

iii) Assume that $k \geq n$. If u(t) is not persistently exciting of order n+k, then R is singular.

Remark: In the not described case, when $k < n$ and u(t) is not persistently exciting of order n+k, nothing general can be stated.

Proof: Consider the equation

$$Rx = 0 \tag{2.5}$$

or equivalently

$$x^T Rx = 0 \tag{2.6}$$

With notations from and use of lemma 2.3 this is written

$$r_\varepsilon(0) = 0 \tag{2.7}$$

Since then $r_\varepsilon(\tau) = 0$ all τ, it follows that (2.7) is equivalent to

$$r_{\varepsilon_1}(0) = 0 \quad \varepsilon_1(t) = A(q^{-1})\varepsilon(t)$$

Now $\varepsilon_1(t) = A(q^{-1})[F(q^{-1})y(t) + G(q^{-1})u(t)] =$

$$= [F(q^{-1})B(q^{-1}) + G(q^{-1})A(q^{-1})]u(t) \equiv H(q^{-1})u(t)$$

The original equation (2.1) is thus transformed into $h^T R_u h = 0$ or

$$R_u h = 0 \tag{2.8}$$

with

$$R_u = \begin{bmatrix} r_u(0)\dots & & r_u(n+k-1) \\ & & \\ & & r_u(0) \end{bmatrix}$$

$$h = \begin{bmatrix} h_1 \\ \\ h_{n+k} \end{bmatrix}$$

Separate two cases.

Case a): Assume that $u(t)$ is persistently exciting of order $n+k$. (2.8) implies $h = 0$ or $H(q^{-1}) \equiv 0$.

If $F(q^{-1}) \not\equiv 0$ it is then concluded that

$$\frac{B(q^{-1})}{A(q^{-1})} = - \frac{G(q^{-1})}{F(q^{-1})} \tag{2.9}$$

where the left hand side is of order n and the right hand side of order $k-1$.

If $k \leq n$ this is a contradiction and $F(q^{-1}) \equiv G(q^{-1}) \equiv 0$, or $x = 0$ is the only solution of (2.5) which proves part i).

If, on the other hand, $k > n$, all solutions of (2.9) are of the form $G(q^{-1}) = -B(q^{-1})\, L(q^{-1})$, $F(q^{-1}) = A(q^{-1})\, L(q^{-1})$ where $L(q^{-1}) = \sum_{i=1}^{k-n} l_i q^{-i}$ is arbitrary. This proves part ii).

The equation $H(q^{-1}) \equiv 0$ can be transformed to a system of linear equations

$$Tx = 0$$

with x as before and T a $(n+k)$ by $2k$ matrix, depending on $a_1, \ldots, a_n, b_1, \ldots, b_n$. More explicitly T is the matrix

$$T = \left[\begin{array}{ccc|ccc}
0 & & 0 & 1 & & 0 \\
b_1 & \ddots & & a_1 & \ddots & \\
 & \ddots & 0 & & \ddots & 1 \\
b_n & & b_1 & a_n & & a_1 \\
 & \ddots & & & \ddots & \\
0 & & & 0 & & \\
 & & b_n & & & a_n
\end{array}\right]$$

From the discussion it is clear that the null space of T $N(T) = \{0\}$

if and only if $k \leqslant n$.

Case b): Assume that $u(t)$ is not persistently exciting of order $n+k$. Then (2.5) is equivalent to $h \in N(R_u)$. Let r be an arbitrary vector in the null space $N(R_u)$. By transforming the equation as before

$$T\,x = r \tag{2.10}$$

If $k > n$ take $r = 0$ and x as (2.1) - (2.4).
If $k = n$, T is a square, invertible matrix and to every $r \neq 0$ there is a non trivial solution of (2.10). This proves part iii).

Q.E.D.

Interpretation: Consider $V = r_\varepsilon(0)$,

$$\varepsilon(t) = F(q^{-1})y(t)+G(q^{-1})u(t),\ F(q^{-1}) = \sum_1^k f_i q^{-i},\ G(q^{-1}) = \sum_1^k g_i q^{-i}$$

The system covariance matrix of order 2k is singular if and only if the minimum of V with respect to $\{f_i\}$ and $\{g_i\}$ is zero.
Loosely speaking the result of the theorem is:
If $k > n$, the filters $F(q^{-1})$ and $G(q^{-1})$ are of higher order than the system and it is possible to get $V=0$.

If $k \leq n$ it is not possible to get $V = 0$ if all modes of the system are excited.

III. MAIN RESULTS

3.1. Introduction

In this chapter the first version of GLS is closer examined. First it is shown (theorem 3.1) that the method can be interpreted as adapting the maximum likelihood technique to this problem. The question of convergence is then reduced to an examination of local maximum points of the likelihood function. It is rather easy to give conditions which guarantee a unique global maximum of the likelihood function (lemma 3.1). As the computations of the GLS method must be carried out on a computer the possible existence of several local maximas is of greater interest. In three theorems it is shown that the number of local maximum points depends on the signal to noise ratio and the order of the model (theorems 3.2, 3.3 and 3.4).

The second version can be interpreted similarily. In the end of this chapter it is shown how to construct examples, where this version of GLS converges to biased estimates.

3.2. Maximum Likelihood Interpretation

In this section it is shown how the GLS method can be interpreted as the maximum likelihood method. Expressions for a corresponding loss function are given in matrix notations and using operators. Finally the limit of this function, as the number of samples tends to infinity, is studied.

Theorem 3.1: Assume that the disturbances are given by

$$v(t) = \frac{1}{C(q^{-1})} e(t) \tag{3.1}$$

e(t) white Gaussian noise. The first version of the GLS method is equivalent to maximizing the likelihood function of this problem by a relaxation method.

Proof:

The probability function of y is given by

$$f(y) = \frac{1}{(2\pi)^{N/2}(\det R)^{1/2}} \exp\left[-\frac{1}{2}(Y-\phi\theta)^T R^{-1}(Y-\phi\theta)\right]$$

(3.1) is written by matrix notations

$$e = Fv$$

$$F = \begin{bmatrix} 1 & & & \\ c_1 & & 0 & \\ & & & \\ c_n & & & \\ & 0 & c_n \ldots\ldots & 1 \end{bmatrix}$$

From (1.12) it follows that

$$R = Evv^T = \left(\frac{1}{\sigma^2} F^T F\right)^{-1}$$

The likelihood function is given by

$$-\log L = \frac{1}{2}(Y-\phi\hat{\theta})^T \frac{1}{\hat{\sigma}^2} F^T F(Y-\phi\hat{\theta}) + \frac{1}{2}\log\det(\hat{F}^{-1}\hat{\sigma}^2(\hat{F}^T)^{-1}) + \frac{N}{2}\log 2\pi \qquad (3.2)$$

Let

$$W(\hat{\theta},F) = \frac{1}{2N}(Y-\phi\hat{\theta})^T \hat{F}^T\hat{F}(Y-\phi\hat{\theta}) \qquad (3.3)$$

so

$$-\log L = \frac{N}{\hat{\sigma}^2} W(\hat{\theta},\hat{F}) + \frac{1}{2}\log(\hat{\sigma}^{2^N}) + \frac{N}{2}\log 2\pi$$

since $\det \hat{F} = 1$

$$\frac{\partial L}{\partial \hat{\sigma}^2} = 0 \text{ implies } -\frac{NW(\hat{\theta},\hat{F})}{(\hat{\sigma}^2)^2} + \frac{N}{2\hat{\sigma}^2} = 0$$

so L is maximized with respect to $\hat{\sigma}^2$ by

$$\hat{\sigma}^2 = 2W(\hat{\theta},\hat{F}) \qquad (3.4)$$

Maximizing L is then equivalent to minimizing $W(\hat{\theta},\hat{F})$. The actual algorithm can be interpreted as a minimization of this function by alternating between

1. Minimize $W(\hat{\theta}_k,\hat{F}_k)$ with respect to $\hat{\theta}_k$

2. Minimize $W(\hat{\theta}_k,\hat{F}_{k+1})$ with respect to $\hat{F}_{k+1}$

which is a relaxation method.

Q.E.D.

Remark 1: Denote the estimate of $a_1,\ldots,a_n,b_1,\ldots,b_n,c_1,\ldots,c_n$ by $\hat{\varphi}_N$ when the record length is N. It follows from [3] and [6] that $\hat{\varphi}_N$ has nice asymptotic properties:

1. $\hat{\varphi}_N$ converges with probability one to the true parameter vector φ as N increases.

2. $\hat{\varphi}_N$ is asymptotic efficient (i.e. has minimal variance).

3. $\hat{\varphi}_N$ is asymptotic normal with the mean value φ and the covariance matrix

$$\frac{2W^{\cdot}}{N}\,W''^{-1}_{\varphi\varphi}$$

Remark 2 $W(\hat{\theta}_k,\hat{F}_k)$ is a decreasing, bounded sequence, which implies convergence. Possible bounded limits must be stationary points of $W(\hat{\theta},\hat{F})$. They cannot be local maximum points. It is shown in Appendix B that saddle points have not to be considered either, since they are not "stable" points. By this concept it is meant that a start of the iteration sufficiently closed to a saddle point will not in general imply convergence to the point. Since the minimization of $W(\hat{\theta},\hat{F})$ has to be carried out on a computer, rounding errors must be introduced in the calculations, and the probability of convergence to a saddle point can for practical cases be regarded as zero. Local minimum points are thus the only "practically possible", bounded limits of $(\hat{\theta}_k,\hat{F}_k)$ as $k\to\infty$.

Remark 3 Note that the convergence of the minimization algorithm is very slow. It is shown in Appendix B that close to a minimum point $\hat{\theta}_k$ will converge linearly.

Remark 4 The second version of GLS can be interpreted in a similar way. Let $W(\hat{\theta},\hat{F})$ be defined from (3.3) and put $\hat{F} = \prod_{i=1}^{\infty} \hat{F}_i$. The iteration procedure is a minimization with different constraints on $\hat{F}$. I.e. in step k, $\hat{F}_1,\ldots,\hat{F}_{k-1}$ are fixed. $W(\hat{\theta}_k,\hat{F})$ is minimized with respect to $\hat{F}_k$. $\hat{F}_{k+1} = \hat{F}_{k+2} = \ldots . I$. This step corresponds to the estimation of the filter $\hat{C}_k(q^{-1})$. From this interpretation it is clear that $W(\hat{\theta},\hat{F})$ is decreased in each step. This fact is shown by straight forward calculations in [18].

From the discussion in 1.3 it is clear that the loss function $W(\hat{\theta},\hat{F})$ can be expressed as

$$W(\hat{a}_1,\dots,\hat{a}_n\ \hat{b}_1,\dots,\hat{b}_n\hat{c}_1,\dots,c_n) = \frac{1}{2N}\sum_{t=1}^{N}\varepsilon^F(t)^2 \tag{3.5}$$

$$\varepsilon^F(t) = \hat{C}(q^{-1})\varepsilon(t) = \varepsilon(t) + \hat{c}_1\varepsilon(t-1)+\dots+\hat{c}_n\varepsilon(t-n) \tag{3.6}$$

$$\varepsilon(t) = \hat{A}(q^{-1})y(t) - \hat{B}(q^{-1})u(t) \tag{3.7}$$

so

$$\varepsilon^F(t) = \hat{C}(q^{-1})\,\frac{\hat{A}(q^{-1})B(q^{-1})-A(q^{-1})\,B\,(q^{-1})}{A(q^{-1})}u(t)\ +$$

$$+\ \frac{\hat{A}(q^{-1})\hat{C}(q^{-1})}{A(q^{-1})}\ v(t) \tag{3.8}$$

Clearly W is a polynomial in $\hat{a}_1\dots\dots\hat{c}_n$ where the coefficients are different sample covariances. An analysis of W and especially the local minimum points of this function must therefore be done in a probabilistic setting. In order to do the analysis reasonable asymptotic theory will be used.

In the following some assumptions are made

i) $u(t) = u_1(t) + G(q^{-1})e_1(t)$

$u_1(t)$ is deterministic, and almost periodic.$G(q^{-1})$ is a stable filter of finite order.
$e_1(t)$ is white noise.

ii) $v(t) = H(q^{-1})e_2(t)$

$H(q^{-1})$ is a stable filter of finite order
$e_2(t)$ is white noise

iii) $e_1(t)$ and $e_2(t)$ ($u(t)$ and $v(t)$) are independent.

Under these assumptions it follows from Theorem 2.1 that W has a limit $V(\hat{a}_1 \ldots \hat{a}_n\ \hat{b}_1 \ldots \hat{b}_n\ \hat{c}_1 \ldots \hat{c}_n)$ with probability one, and that

$$V(\hat{a}_1 \ldots \hat{c}_n) = V_1(\hat{a}_1 \ldots \hat{c}_n) + V_2(\hat{a}_1 \ldots \hat{c}_n) \tag{3.9}$$

$$V_i(\hat{a}_1 \ldots \hat{c}_n) = \frac{1}{2} E\varepsilon_i^F(t)^2 \tag{3.10}$$

$$\varepsilon_1^F(t) = \hat{C}(q^{-1}) \frac{\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1})}{A(q^{-1})} u(t) \tag{3.11}$$

$$\varepsilon_2^F(t) = \frac{\hat{A}(q^{-1})\hat{C}(q^{-1})H(q^{-1})}{A(q^{-1})} e(t) \tag{3.12}$$

The notation $Eu_1^2(t)$ denotes

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} u_1^2(t).$$

It is the purpose of Sections 3.3 - 3.7 closer to examine the loss function V. The main interest will be an investigation when the loss function has a unique local minimum.

In order to simplify the analysis a bit only "interesting" values of the parameter estimates will be considered.

In many cases the following compact set in the parameter space will be reasonable:

i) $\hat{A}(z)$ has all zeros inside the circle $|z| \leq r<1$.

ii) $C(z)$ " " " " " $|z| \leq r<1$.

iii) $\hat{b}_i$ bounded.
r close to 1.

This restriction is well justified by physical reasons.

i) means that a stable model is required,

ii) is motivated by the representation theorem [2] and finite variance of the output,

iii) must be fulfilled if the model has finite gain.

3.3. Global properties of the loss function.

This section contains some simple considerations concerning the global minimum of the loss function in the special case

$$v(t) = \frac{1}{C(q^{-1})} e(t) \tag{3.13}$$

Lemma 3.1: Consider the loss function (3.9) with v(t) as (3.13). Denote the order of the model by m and the order of the system by n. Assume m ⩾ n.

i) Global minimum points are the solution of

$$\begin{cases} \hat{A}(q^{-1})\hat{C}(q^{-1}) = A(q^{-1})C(q^{-1}) \\ \\ \varepsilon_1^F(t) = \hat{C}(q^{-1}) \dfrac{\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1})}{A(q^{-1})} u(t) = 0 \\ \\ \text{with probability one.} \end{cases} \tag{3.14}$$

ii)
$$\begin{matrix} \hat{a}_i = a_i & i = 1, \dots n & \hat{a}_i = 0 & i = n+1, \dots m \\ \hat{b}_i = b_i & i = 1, \dots n & \hat{b}_i = 0 & i = n+1, \dots m \\ \hat{c}_i = c_i & i = 1, \dots n & \hat{c}_i = 0 & i = n+1, \dots m \end{matrix} \tag{3.15}$$

is always a global minimum point.

iii) If $u(t)$ is persistently exciting of order $n+m$, $m=n$, and the system is controllable, then (3.15) is the unique global minimum point.

iv) If $u(t)$ is not persistently exciting of order $n+m$, $m=n$, there may exist other global minimum points than (3.15).

v) If $u(t)$ is persistently exciting of order $n+m$, $m>n$, there are in general several global minimum points. These points are equivalent in the sense that they all satisfy

$$\frac{\hat{B}(q^{-1})}{\hat{A}(q^{-1})} = \frac{B(q^{-1})}{A(q^{-1})} \tag{3.16}$$

$$\frac{1}{\hat{A}(q^{-1})\hat{C}(q^{-1})} = \frac{1}{A(q^{-1})C(q^{-1})} \tag{3.17}$$

Proof: Clearly $\inf V_1 = 0$. Further $\inf V_2 = \frac{1}{2} Ee^2(t)$. To realize that define

$$G(q^{-1}) = \frac{\hat{A}(q^{-1})\hat{C}(q^{-1})}{A(q^{-1})C(q^{-1})} = 1 + \sum_{i=1}^{\infty} g_i q^{-i}$$

Then

$$V_2 = \frac{1}{2} Ee^2(t)\left[1 + \sum_{i=1}^{\infty} g_i^2\right]$$

and $\inf V_2 = \frac{1}{2} Ee^2(t)$ for $g_i = 0$, $i = 1, \ldots$

i) The equations

$$V_1 = \inf V_1$$
$$V_2 = \inf V_2$$

have the solutions (3.14).

ii) The assertion follows directly from i).

iii) From Lemma 2.2 it is concluded that

$$\hat{A}(q^{-1})B(q^{-1}) \equiv A(q^{-1})\hat{B}(q^{-1})$$

and by arguments as in the proof of Theorem 2.1 the assertion follows.

iv) An example for a first order system with $a \neq c$

$$u(t) = \lambda^t, \quad \lambda = \pm 1$$

$$\hat{a} = c, \qquad \hat{b} = b\,\frac{1 + c\lambda}{1 + a\lambda}, \quad \hat{c} = a$$

v) Lemma 2.2 implies

$$\hat{A}(q^{-1})B(q^{-1}) \equiv A(q^{-1})\hat{B}(q^{-1})$$

so (3.16) is proved. (3.17) follows from (3.14). In general the factor in common between $\hat{A}(q^{-1})$ and $\hat{B}(q^{-1})$ can be chosen in several ways to satisfy (3.17).

Q.E.D.

Remark: The assumption that (3.13) holds is essential for the result.

3.4. Estimates at high signal to noise ratios. Models of correct order.

In this section a theorem of uniqueness is given and discussed. The essential part of the proof is found in Appendix C as a series of lemmas.

Theorem 3.2: Let the system of order n

$$A(q^{-1})y(t) = B(q^{-1})u(t) + v(t), \quad v(t) = H(q^{-1})e(t) \tag{3.18}$$

be controllable and the input u(t) persistently exciting of order 2n. Assume that the order of the model is n. Consider parameter estimates in Ω, an arbitrary compact set.

Then there is a constant S_o such that if $S_o \leqslant S < \infty$ then the loss function (3.9) has exactly one stationary point in Ω. This point is a local minimum and satisfies

$$\begin{cases} \hat{a}_i = a_i + O(1/S) & i = 1 \dots n \\ \hat{b}_i = b_i + O(1/S) & i = 1 \dots n \\ \hat{c}_i = \bar{c}_i + O(1/S) & i = 1 \dots n \end{cases} \tag{3.19}$$

where $\bar{C}(q^{-1}) = 1 + \bar{c}_1 q^{-1} + \dots + \bar{c}_n q^{-n}$ and $(\bar{c}_1, \dots \bar{c}_n)$ is the minimum point of

$$E\left[\hat{C}(q^{-1})v(t)\right]^2 \tag{3.20}$$

Proof: Introduce the vectors x and y by

$$x = \begin{bmatrix} \hat{a}_1 - a_1 \\ \vdots \\ \hat{a}_n - a_n \\ \hat{b}_1 - b_1 \\ \vdots \\ \hat{b}_n - b_n \end{bmatrix} \qquad y = \begin{bmatrix} \hat{c}_1 \\ \vdots \\ \hat{c}_n \end{bmatrix}$$

Then the loss function (3.9) can be written

$$V(x,y) = \frac{1}{2} x^T P(y)x + \varepsilon h(x,y)$$

with P(y) as the covariance matrix of the system

$$A(q^{-1})y^F(t) = - B(q^{-1})u^F(t)$$

$$u^F(t) = \hat{C}(q^{-1})u(t)$$

This fact follows from Lemma 2.3. From corr of Lemma 2.1 and Theorem 2.2 follows that P(y) is non singular for all y. Further the loss function is assumed to be scaled so that ε denotes the quantity 1/S.

The function $h(0,y) = 2\ E[\hat{C}(q^{-1})v(t)]^2$ is quadratic in y. It has a unique minimum point y_o, which fulfils $h''_{yy}(0,y_o)$ positive definite. Invoking Theorem C.1 the proof is finished.

Q.E.D.

What sense have the different assumptions?

i) The restriction on the input signal is very natural. This condition is necessary for the result (Lemma 3.1).

ii) The study of only parameter estimates in Ω is motivated before.

iii) The restriction on the signal to noise ratio is crucial as is shown in Theorem 3.3.

iv) The assumption of controllability is essential. If the system is non controllable, there is a factor in common between $A(q^{-1})$ and $B(q^{-1})$. Equation (3.18) can be divided by this factor, obtaining a controllable system of lower order than the original and with another correlation of the noise. If the system is not controllable, it is thus equivalent to regard the order of the model as higher than the order of the (controllable part of the) system. This situation is treated in Section 3.7, where it is shown that non controllable systems in general will give no unique local minimum.

3.5. Estimates at low signal to noise ratios.

This section deals with the case of low signal to noise ratios. It turns out that a possible property of the noise plays an essential role for non uniqueness.

Definition 3.2. The noise $v(t) = H(q^{-1})e(t)$ fulfils the "noise condition" (NC) if there exist at least two different pairs of polynomials $\hat{A}_1(q^{-1})$, $\hat{C}_1(q^{-1})$ and $\hat{A}_2(q^{-1})$, $\hat{C}_2(q^{-1})$, such that

$$V_2(\hat{a}_1 \dots \hat{a}_n,\ \hat{c}_1 \dots \hat{c}_n) = E\left[\frac{\hat{A}(q^{-1})\hat{C}(q^{-1})H(q^{-1})}{A(q^{-1})}\, e(t)\right]^2 \qquad (3.9)$$

has a local minimum point with a positive definite matrix of second order derivatives in $(\hat{a}_{11} \dots \hat{a}_{1n},\ \hat{c}_{11} \dots \hat{c}_{1n})$ and $(\hat{a}_{21} \dots \hat{a}_{2n},\ \hat{c}_{21} \dots \hat{c}_{2n})$.

Remark:

$$w(t) = \frac{H(q^{-1})}{A(q^{-1})} e(t)$$

is the measurement noise if all noise of the process is interpreted as measurement noise.

Corr 1: v(t) fulfils (NC) if there exists a minimum point with V_2'' positive definite, $\hat{A}(q^{-1}) \not\equiv \hat{C}(q^{-1})$.

Proof: Take $\hat{A}_2 = \hat{C}$, $\hat{C}_2 = \hat{A}$. By symmetry this corresponds to another point satisfying the predescribed conditions.

Corr 2: If

$$v(t) = \frac{1}{C(q^{-1})} e(t)$$

it is sufficient that there is a factorization of $A(q^{-1})$ and $C(q^{-1})$ such that $A(q^{-1})C(q^{-1}) = A_1(q^{-1})C_1(q^{-1})$ where $A_1(q^{-1})$ and $C_1(q^{-1})$ have no factors in common.

Proof: $\hat{A}_1(q^{-1}) = A_1(q^{-1})$, $\hat{C}_1(q^{-1}) = C_1(q^{-1})$ and $\hat{A}_2(q^{-1}) = C_1(q^{-1})$, $\hat{C}_2(q^{-1}) = A_1(q^{-1})$ define two different (local and global) minimum points. The matrix of second order derivatives is given by

$$\frac{\partial^2 V_2}{\partial \hat{a}_i \, \partial \hat{a}_j} = 2E\left[\left(q^{-i}\hat{C}(q^{-1})v(t)\right)\left(q^{-j}\hat{C}(q^{-1})v(t)\right)\right]$$

$$\frac{\partial^2 V_2}{\partial \hat{a}_i \, \partial \hat{c}_j} = 2E\left[\left(q^{-i}\hat{C}(q^{-1})v(t)\right)\left(q^{-j}\hat{A}(q^{-1})v(t)\right)\right] +$$

$$+ \; 2E\left[\left(q^{-i-j}v(t)\right)\left(\hat{A}(q^{-1})\hat{C}(q^{-1})v(t)\right)\right]$$

$$\frac{\partial^2 V_2}{\partial \hat{a}_i \partial \hat{c}_j} = 2E\left[\left[q^{-i}\hat{A}(q^{-1})v(t)\right]\left[q^{-j}\hat{A}(q^{-1})v(t)\right]\right]$$

With $\hat{A}(q^{-1})\hat{C}(q^{-1}) = A(q^{-1})C(q^{-1})$ the second term of $\partial^2 V_2/\partial\hat{a}_i\partial\hat{c}_j$ vanishes and $\frac{1}{2}\ V_2''$ becomes the system covariance matrix of

$$\hat{A}(q^{-1})y(t) = \hat{C}(q^{-1})u(t), \quad u(t) = \hat{A}(q^{-1})v(t)$$

From Theorem 2.1 it follows that V_2'' is positive definite.

Q.E.D.

It would be valuable to know, when (NC) is fulfilled in general. However, (NC) is depending on the orders of $\hat{A}(q^{-1})$ and $\hat{C}(q^{-1})$ and the correlation of the noise. Some results for the simple case of first order models are given in Section 3.6.

The concept of (NC) is now used in a theorem of non uniqueness.

Theorem 3.3. Assume that the noise v(t) fulfils (NC). Then there is a number $S_1 > 0$ such that $0 < S \leqslant S_1$ implies that the loss function V (3.9) has more than one local minimum.

Remark: The result of the theorem holds only for sufficiently small values of the signal to noise ratio. Simulations show, however, see Chapter 4, that the result may be true also for reasonable values of S.

Proof: It will be shown that V has (at least) two local minimum points satisfying

$$\begin{cases} \hat{A}(q^{-1}) = A_i(q^{-1}) + O(S) \\ \hat{C}(q^{-1}) = C_i(q^{-1}) + O(S) \end{cases} \qquad i = 1,\ 2 \tag{3.21}$$

It follows from the proof of Theorem 3.2 that the equations

$$\frac{\partial V}{\partial \hat{b}_i} = 0 \qquad i = 1, \ldots, n$$

are a system of linear equations in the unknown parameters. The system has always a unique solution, depending on $\hat{a}_i$ and $\hat{c}_i$ but not on S.

Put this solution into the remaining equations.

$$\begin{cases} \dfrac{\partial V}{\partial \hat{a}_i} = 0 \qquad i = 1, \ldots, n \\ \\ \dfrac{\partial V}{\partial \hat{c}_i} = 0 \qquad i = 1, \ldots, n \end{cases} \tag{3.22}$$

(3.22) is now written in the form

$$0 = V_2'(x) + S\bar{V}_1'(x) \tag{3.23}$$

where it has been assumed that $\sigma^2 = Ee^2(t) = 1$. x denotes the vector $[\hat{a}_1 \ldots \hat{a}_n, \hat{c}_1 \ldots \hat{c}_n]^T$.

(NC) implies the existence of two points x_1 and x_2 satisfying

$$\begin{cases} V_2'(x_i) = 0 \\ V_2''(x_i) \text{ positive definite} \end{cases} \qquad i = 1,\ 2 \tag{3.24}$$

From Lemma C.3 it follows that the solutions (3.21) exist.

When the variables are ordered as

$$[\hat{a}_1 \ldots \hat{a}_n \; \hat{c}_1 \ldots \hat{c}_n \; \hat{b}_1 \ldots \hat{b}_n]$$

the matrix of second order derivatives will be

$$V'' = \begin{bmatrix} V_2'(x_i + O(S) & O(S) \\ & \\ O(S) & SP \end{bmatrix}$$

where P is a positive definite matrix. From Lemma B.5 it follows that V" is positive definite and that the obtained solutions of V' = 0 are local minimum points.

Q.E.D.

Bohlin [5] has given results, which can be used to test if an estimate is the true maximum likelihood estimate. The test quantity involves sample covariances of $\varepsilon(t)$ and u(t). If, however, the noise level is high this method cannot be used successfully in the case described here. The minimum points of the loss function will give residuals $\varepsilon_1(t)$ and $\varepsilon_2(t)$ satisfying $\varepsilon_1(t) - \varepsilon_2(t) = O(S)$ so also all possible test quantities will differ just a little if S is small.

3.6. Analysis of the "noise condition" (NC) for first order models.

The noise condition (NC) will be closer analysed for first order models in this section. In this case the loss function (3.9) reduces to

$$V_2(\hat{a}_1,\hat{c}) = [1 + (\hat{a}+\hat{c})^2 + \hat{a}^2\hat{c}^2]r_o + \\ + [2(\hat{a}+\hat{c})(1+\hat{a}\hat{c})]r_1 + [2\hat{a}\hat{c}]r_2 \qquad (3.25)$$

where

$$r_\tau = r_w(\tau)$$

$$w(t) = \frac{H(q^{-1})}{A(q^{-1})} e(t)$$

An analysis of this function is made in

Lemma 3.2. For models of order one (NC) is fulfilled if and only if

$$D^* = r_1^2(r_2-r_o)^2 - 4(r_o^2-r_1^2)(r_1^2-r_or_2) > 0 \qquad (3.26)$$

Proof: See Appendix D.

The following examples illustrate the fact that the noise condition depends on the covariance function of the measurement noise w(t).

Example 1:

$$w(t) = \frac{1}{(1+aq^{-1})(1+cq^{-1})} e(t)$$

(NC) is fulfilled if and only if $a \neq c$ (Corr 2 of Def. 3.1).

Example 2:

$$r_2 = 0$$

Then $D^* > 0$ if and only if

$$|r_1| > \frac{\sqrt{3}}{2} r_o$$

For the special structure $w(t) = (1 + cq^{-1})e(t)$ this is never fulfilled.

Example 3:

$$r_1 = 0$$

Then $D^* = 4 r_o^3 r_2$ and the sign of D^* is equal to the sign of r_2. For the special structure $w(t) = (1 + \gamma q^{-2}) \cdot e(t)$. $D^* > 0$ if and only if $\gamma > 0$, i.e. $(z^2 + \gamma)$ has zeros on the imaginary axis.

Example 4:

$$w(t) = \frac{1 + cq^{-1}}{1 + aq^{-1}} e(t)$$

Up to second order terms in a and c D^* is given by

$$D^* = (a - c)(3a + 7c) + \ldots$$

This expression indicates that a rather involved relation between a and c determines if (NC) is fulfilled or not.

3.7. Estimates at high signal to noise ratios. Models of too high an order.

Since the true order of a system seldom is known in practice, it is valuable to know what will happen if the model has higher order than the system. In this section it is shown how the result for models of correct order (Theorem 3.2) can be generalized.

The result of Theorem 3.2 can be described as follows. Neglecting terms $O(1/S)$ the (unique) minimum satisfies

i) it minimizes $V_1(\hat{a}_1 \ldots \hat{b}_i \ldots \hat{c}_n)$,

ii) with the remaining degrees of freedom it minimizes $V_2(\hat{a}_1 \ldots \hat{c}_n)$.

If the order of the model is greater than the order of the system it will turn out that there may be more than one minimum point, but the characterization above still applies if local minimum points are concerned under part ii).

Theorem 3.4: Let the system

$$A(q^{-1})y(t) = B(q^{-1})u(t) + v(t), \quad v(t) = H(q^{-1})e(t) \quad (3.27)$$

be controllable and of order n.

Assume that the order of the model is $n+k$, $k > 0$ and that $u(t)$ is persistently exciting of order $2n+k$ Consider parameter estimates in Ω, an arbitrary compact set. Then there is a constant S_o such that if $S_o \leq S < \infty$.

i) All local minimum points of the loss function (3.9) fulfil

$$\hat{A}(q^{-1}) = A(q^{-1})L(q^{-1}) + o(1), \quad S \to \infty \tag{3.28}$$

$$\hat{B}(q^{-1}) = B(q^{-1})L(q^{-1}) + o(1), \quad S \to \infty \tag{3.29}$$

where $L(q^{-1}) = 1 + \ell_1 q^{-1} + \ldots + \ell_k q^{-k}$.

Further $L(q^{-1})$ and $\hat{C}(q^{-1})$ fulfil

$$L(q^{-1}) = \bar{L}(q^{-1}) + o(1), \quad S \to \infty \tag{3.30}$$

$$\hat{C}(q^{-1}) = \bar{C}(q^{-1}) + o(1), \quad S \to \infty \tag{3.31}$$

where $(\bar{\ell}_1, \ldots, \bar{\ell}_k, \bar{c}_1, \ldots, \bar{c}_{n+k})$ is a stationary point of

$$V_3(\ell_1, \ldots, \ell_k, c_1, \ldots, c_{n+k}) =$$

$$= E[L(q^{-1})\hat{C}(q^{-1})v(t)]^2 \tag{3.32}$$

The matrix of second order derivatives of V_3 in $(\bar{\ell}_1, \ldots, \bar{c}_{n+k})$ must be positive definite or positive semidefinite.

ii) If the matrix of second order derivatives of V_3 in $(\bar{\ell}_1, \ldots, \bar{c}_{n+k})$ is positive definite, then there exists a unique local minimum point of the form (3.28) - (3.31) and the terms o(1) can be replaced by O(1/S). Further the matrix V" is positive definite in this point.

Proof: See Appendix E.

Remark 1: The number of stationary points of V_3 and the number of local minimum points of V are coupled to the condition (NC) introduced in Section 3.5.

Remark 2: All possible local minimum points have the property

$$\frac{\hat{B}(q^{-1})}{\hat{A}(q^{-1})} = \frac{B(q^{-1})}{A(q^{-1})} + o(1), \qquad S \to \infty$$

Remark 3: If V_3'' is singular in $(\bar{\ell}_1, \ldots, \bar{c}_{n+k})$ nothing general can be stated. In the special case

$$v(t) = \frac{1}{C(q^{-1})} e(t)$$

all points $(\ell_1, \ldots, \ell_k, \hat{c}_1, \ldots, \hat{c}_{n+k})$ satisfying

$$L(q^{-1})\hat{C}(q^{-1}) \equiv C(q^{-1}) \tag{3.33}$$

are global minimum points of V_3 and for some of them V_3'' is singular. This follows from the proof of Corr 2 of Def. 3.2. However, from Lemma 3.1, part i), it follows that the points satisfying (3.33) correspond to global minimum points of the loss function.

3.8. Counter examples to convergence of the second version of GLS.

In this section an example illustrating the possible behaviour of the second version of GLS is described. The question of convergence of this version under suitable conditions cannot be answered easily, and it has not been studied by the author.

The following case will be taken into consideration. The system and the model are both of first order. The iteration is started with the LS estimate of a and b. Conditions for convergence in the next step are examined. If the estimated operator $\hat{C}(q^{-1}) \equiv 1$ then the following estimation of a and b will give the same result as before.

The interesting equations are thus:

$$\hat{c} = -\frac{r_{\varepsilon}(1)}{r_{\varepsilon}(0)} = 0 \tag{3.34}$$

$$\varepsilon(t) = (1 + \hat{a}q^{-1})y(t) - \hat{b}q^{-1}u(t) \tag{3.35}$$

$$\begin{bmatrix} r_y(0) & -r_{yu}(0) \\ -r_{yu}(0) & r_u(0) \end{bmatrix} \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} -r_y(1) \\ r_{yu}(1) \end{bmatrix} \tag{3.36}$$

<u>Example</u>: Consider the system

$$(1 + aq^{-1})y(t) = u(t) + v(t), \qquad v(t) = \frac{1}{1 + cq^{-1}}\, e(t)$$

where u(t) is white noise. There is a number $S_o > 0$ such that if $0 < S \leqslant S_o$ then (3.34) has two solutions w.r.t.a, which satisfy

$$a = O(S), \qquad S \to 0$$

$$a = -c + O(S), \qquad S \to 0$$

In Appendix F the existence of these solutions are proved.

Note that in this example of first order systems, only systems with a special value of a and a low signal to noise ratio will converge to biased estimates. Nevertheless, the examples indicate that the method may yield "wrong" results. If the iterations procedure is studied in more steps, several more cases of convergence to uncorrect estimates may be detected.

If there is no restriction on the input signal, other examples can be constructed. For example, if the input signal is not persistently exciting of order 2, there are values of a, independent of S, such that the system in the example above will yield biased estimates.

IV. NUMERICAL ILLUSTRATION.

4.1. Introduction.

The theory of the GLS method in Chapter 3 requires an infinite number of data. For practical purposes it is interesting to know if the result holds with "good approximation" for a finite number of data.

The loss function (3.9) is a polynomial in the variables $\hat{a}_1, \ldots, \hat{a}_n, \hat{b}_1, \ldots, \hat{b}_n, \hat{c}_1, \ldots, \hat{c}_n$. The coefficients are different sample covariances, which converge with probability one to the corresponding covariances, as $N \to \infty$. A sufficiently small deviation of the coefficients from their limits can only move the minimum points a little bit, but the probability for a drastical change of the character of the loss function is very small.

This means that for a "sufficiently large" number of data the results of Chapter 3 will hold with probability close to one. However, it is not practically possible to analyze what sufficiently large exactly means.

In order to examine the situation of a finite number of data simulations were used. These simulations are illustrating the results of Chapter 3 as well.

The simulations were carried out on a UNIVAC 1108. A description of the used programs is given in Appendix G.

The results of the simulations are presented in the next sections.

All the simulated systems were generated by the equation

$$A(q^{-1})y(t) = B(q^{-1})u(t) + v(t)$$

$$v(t) = H(q^{-1})e(t)$$

The number of samples were 500 in all cases and the input signal was a PRBS with amplitude 1.0.

4.2. Illustration of Theorem 3.2.

These examples are intended to demonstrate that when the conditions of Theorem 3.2 are fulfilled there is a solution, which satisfies

$$\hat{a}_i \simeq a_i$$

$$\hat{b}_i \simeq b_i$$

$$\hat{c}_i \simeq \bar{c}_i$$

where $\bar{C}(q^{-1})$ corresponds to the minimum point of

$$E[\hat{C}(q^{-1})v(t)]^2$$

The following systems were studied.

Table 4.1 - Generated systems.

System	a_1, (a_2)	b_1, (b_2)	$H(q^{-1})$	$Ee^2(t)$
S1	-0.8	1.0	$\frac{1}{1 + 0.7q^{-1}}$	1.0
S2	-0.8	1.0	$(1 + 0.7q^{-1})$	0.01
S3	-0.8	1.0	$(1-1.0q^{-1}+0.2q^{-2})$	0.01
S4	-1.5 0.7	1.0 0.5	$(1 + 0.7q^{-1})$	0.01

The results of the identifications are given in Table 4.2. The iterations were started with the LS estimation of the a_i and the b_i parameters. The results are very well in accordance with the expectations.

Table 4.2 - Identification results.

System	$\hat{a}_1$, $(\hat{a}_2)$	$\hat{b}_1$, $(\hat{b}_2)$	$\hat{c}_1$, $(\hat{c}_2)$	$\bar{c}_1$, $(\bar{c}_2)$
S1	-0.804	1.010	0.697	0.7
S2	-0.803	1.000	-0.449	-0.469
	-0.803	1.003	-0.607 0.352	-0.603 0.284
S3	-0.799	1.005	0.555	0.589
S4	-1.505 0.704	1.001 0.498	-0.444	-0.469

4.3. Illustration of Theorem 3.3.

For the following systems 500 samples were generated with a PRBS as input signal. The iterations were started with expected values of $\hat{c}_i$.

The system S7 requires a comment. The equation

$$V' = V'(\hat{\theta},\sigma) = 0 \qquad (4.1)$$

where $\hat{\theta}^T = [\hat{a}_1 \ldots \hat{b}_i \ldots \hat{c}_n]$ and $\sigma^2 = Ee^2(t)$ was solved (using analytic expressions for the covariances) with successively decreasing values of the parameter σ. A change $d\sigma$ of σ causes a change in $\hat{\theta}$ approximately

$$d\hat{\theta} = - V''(\hat{\theta},\sigma)^{-1} \frac{\partial}{\partial\sigma} V'(\hat{\theta},\sigma)d\sigma$$

Starting with this new value of $\hat{\theta}$ (4.1) was solved with respect to $\hat{\theta}$ by Newton-Raphson technique. This procedure of computing solutions for different, decreasing values of σ stops when V'' is not positive definite or when $\hat{\theta} \approx \theta_o$ is obtained as solution.

Table 4.3 - Generated systems.

System	a_1	b_1	$H(q^{-1})$	$Ee^2(t)$
S5	-0.8	1.0	$\frac{1}{1 - 0.2q^{-1}}$	100.0
S6	0.0	1.0	$(1 + 0.7q^{-2})$	100.0
S7 (=S1)	-0.8	1.0	$\frac{1}{1 + 0.7q^{-1}}$	1.0

The results presented in Table 4.4 coincide with the predicted values. The last system shows that it is not necessary that the noise has unrealistic high variance for Theorem 3.3 to hold. The expected value of $\hat{b}_1$ is computed from the equation

$$\frac{\partial}{\partial \hat{b}_1} V(\hat{a}_1, \hat{b}_1, \hat{c}_1) = 0$$

where the values of $\hat{a}_1$ and $\hat{c}_1$ are inserted.

Table 4.4 - Identification results.

System	$\hat{a}_1$	$\hat{b}_1$	$\hat{c}_1$	Expected values of $\hat{a}_1$	 $\hat{b}_1$	 $\hat{c}_1$
S5	-0.774	1.051	-0.233	-0.8	1.0	-0.2
	-0.243	0.740	-0.767	-0.2	1.0	-0.8
S6	-0.676	0.705	0.676	-0.69	0.68	0.69
	0.674	0.882	-0.678	0.69	0.68	-0.69
S7	-0.804	1.010	0.697	-0.8	1.0	0.7
	0.327	0.461	-0.771	0.35	0.44	-0.81

4.4. Illustration of Theorem 3.4.

The illustration of Theorem 3.4 has turned out for the author to be more difficult than the previous examples. The reason for this difficulty is probably that the properties (as existence of several minimum points) of the loss function are rather sensitive for the number of data and the realization. This fact is also the reason why the examples in this section require more iterations for convergence.

Analogously to the previous examples the iterations were started with the expected values of the $\hat{c}_i$ parameters.

Table 4.5 - Generated systems.

System	a_1	b_1	$H(q^{-1})$	$Ee^2(t)$
S8	-0.8	1.0	$(1 + 0.8q^{-2})$	0.01
S9	-0.4	1.0	$\dfrac{1}{(1 - 0.8q^{-1})(1 + 0.8q^{-1})}$	1.0

The result of the identifications (see Table 4.6) are well coinciding with the theory.

Table 4.6 - Identification results.

System	From simulation $\hat{a}_1$	$\hat{a}_2$	$\hat{b}_1$	$\hat{b}_2$	$\hat{c}_1$	$\hat{c}_2$
S8	-0.94	0.11	1.00	-0.14	0.11	-0.46
	-0.06	-0.60	1.00	0.74	-0.78	0.11
	-1.39	0.47	1.00	-0.59	0.55	-0.15
S9	-0.45	0.06	0.97	-0.04	0.03	-0.66
	0.40	-0.35	1.00	0.79	-0.82	0.05
	-1.11	0.25	1.01	-0.69	0.69	-0.11

Expected values

S8	-0.80	0.00	1.00	0.00	0.00	-0.45
	-0.13	-0.54	1.00	0.67	-0.67	0.00
	-1.47	0.54	1.00	-0.67	0.67	0.00
S9	-0.40	0.00	1.00	0.00	0.00	-0.64
	0.40	-0.32	1.00	0.80	-0.80	0.00
	-1.20	0.32	1.00	-0.80	0.80	0.00

4.5. Illustration of Section 3.8.

The following examples illustrate that the second version of GLS can converge to "wrong" values of the estimates. The third example, System S12, is constructed in a way similar to System S7. Of course, there is another equation to be solved. (More exactly $\hat{c}(a,c_1\sigma) = 0$ is solved with respect to a for decreasing values of the parameter σ and with fixed value of the parameter c.)

Table 4.7 - Generated systems.

System	a_1	b_1	$H(q^{-1})$	$Ee^2(t)$
S10	-0.5	1.0	$\frac{1}{1 + 0.5q^{-1}}$	100.0
S11	0.0	1.0	$\frac{1}{1 - 0.8q^{-1}}$	100.0
S12	-0.7	1.0	$\frac{1}{1 + 0.9q^{-1}}$	1.2

The results of the identifications, given in Table 4.8, confirm the theory. $\sigma_{\hat{c}_1}$ denotes the estimated standard deviation of $\hat{c}_1$. The PRBS which is used as input signal is with "good approximation" white noise.

Table 4.8 - Identification results.

System	$\hat{a}_1$	$\hat{b}_1$	$\hat{c}_1$	$\sigma_{\hat{c}_1}$	Exp. values of $\hat{a}_1$	 $\hat{b}_1$	 $\hat{c}_1$
S10	-0.01	0.98	0.007	0.045	0.0	1.0	0.0
S11	-0.79	1.08	-0.026	0.045	-0.8	1.0	0.0
S12	0.16	0.94	0.043	0.045	0.09	1.0	0.0

V. EXAMPLES OF LACK OF UNIQUENESS FOR INDUSTRIAL DATA.

5.1. Introduction.

In this chapter identification results using the GLS method of real data are presented. The main purpose of the identifications was to investigate the possible existence of more than one minimum point of the loss function. A straight forward application of a test of order [3] would in general result in more complex models. However, the orders of models in the presented cases are not unreasonable.

The results of the identifications are compared with models obtained with the "ordinary" maximum likelihood model

$$\hat{A}(q^{-1})y(t) = \hat{B}(q^{-1})u(t) + \hat{C}(q^{-1})e(t) \qquad (5.1)$$

It is to be noted that for a "wrong" minimum point the covariance matrix

$$\frac{2V}{N} V''^{-1}$$

of the parameter estimates has dubious meaning.

5.2. Identification of dynamics of a heat rod process.

The system is a copper rod, which acts as a one dimensional heat diffusion process. The system is located at Div. of Automatic Control, Lund Institute of Technology. Identification results using the ML model (5.1) as well as a short description of the process are given in [14]. (The data used here is Serie S1, output $x = 3\ell/4$.)

The test quantity for comparing models of orders 4 and 5, [3], is F(862,3) and has the value 109. Since the ML identification [14] indicates a model of order 4 as reasonable, this order was considered in spite of the great value of the test quantity.

The loss function turns out to have (at least) two minimum points for fourth order models. The results are presented in Tables 5.1 - 5.2 and Figures 5.1 - 5.4.

The theoretical value of the static gain is 0.25, which indicates that model 1 is the most correct one.

In Figures 5.1, 5.2 the following signals are plotted:

1. the input $u(t)$,
2. the output $y(t)$,
3. the model output $y_m(t) = \dfrac{\hat{B}(q^{-1})}{\hat{A}(q^{-1})}\, u(t)$
4. the model error $e_m(t) = y(t) - y_m(t)$
5. the residuals $\varepsilon(t)$.

In Figures 5.3, 5.4 normalized covarious functions are plotted. The criterion by Bohlin [5] can be formulated as: the estimate is true if and only if

$r_{\varepsilon}(\tau) = 0 \qquad \tau > 0$

$r_{\varepsilon u}(\tau) = E\varepsilon(t)u(t+\tau) = 0 \quad \text{all } \tau$

The second condition can also be written as

$r_{e_m}u(\tau) = 0 \quad \text{all } \tau$

Discussion of the results:

Already from the values of the static gain it can be expected that model 1 is superior to model 2. This fact is very much confirmed by the plotted signals.

A comparison with plots of the ML model (see [14]) shows little difference between that model and model 1.

In the model 2 the output is "interpreted" as mainly due to noise.

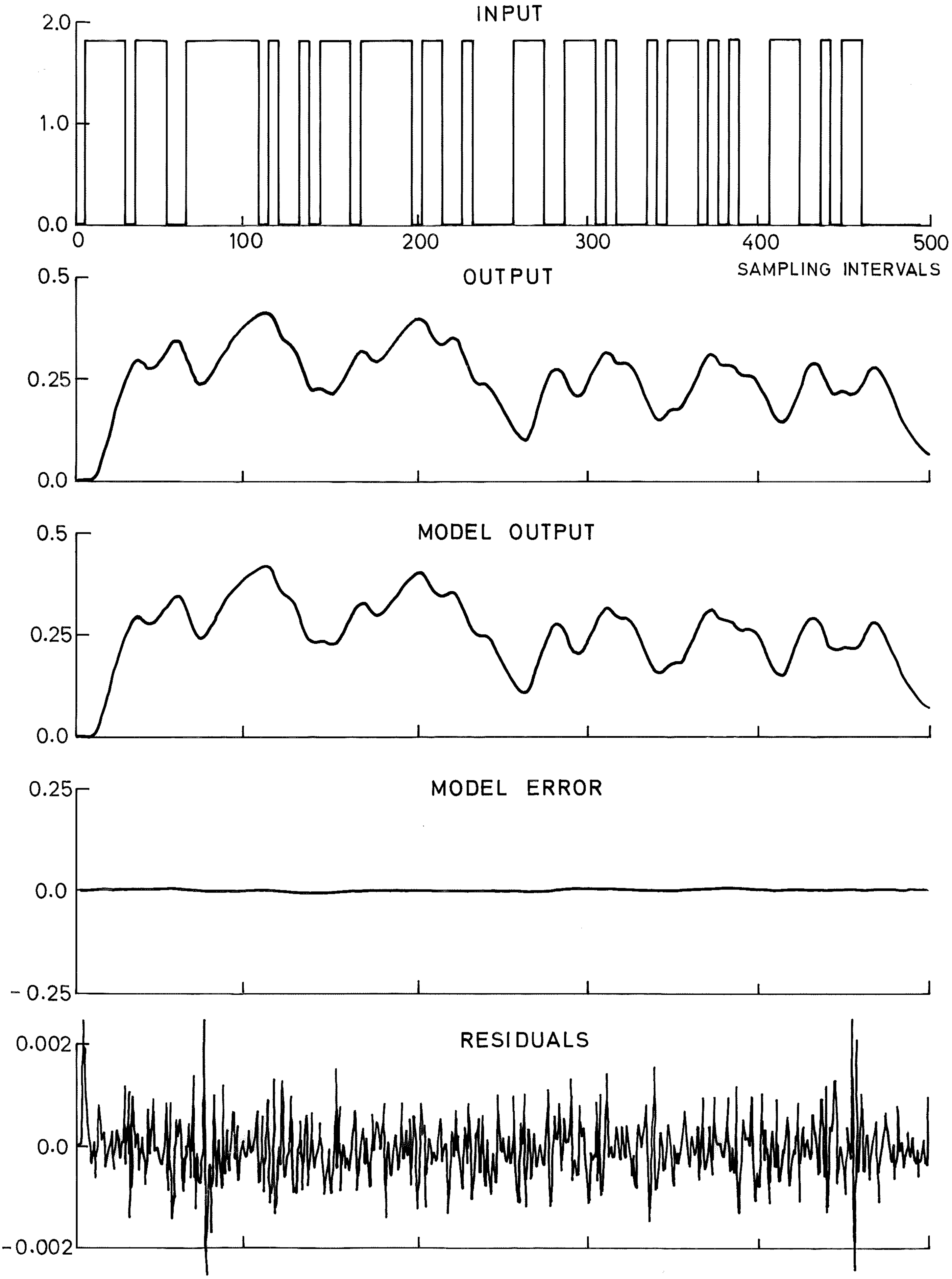
From Figure 5.3 and 5.4 it is seen that the residuals are not white in any of the two models. The input signal and the residuals are considerably more correlated for the second model.

	Model 1	Model 2	Corresponding model in [14]
$\hat{a}_1$	-2.4307±0.0424	-1.2374±0.0510	-2.9563±0.0017
$\hat{a}_2$	1.8776±0.1118	0.6056±0.0888	3.2694±0.0049
$\hat{a}_3$	-0.37271±0.0999	-0.5161±0.0723	-1.6134±0.0047
$\hat{a}_4$	-0.071016±0.0303	0.3525±0.0332	0.3025±0.0015
$\hat{b}_1 \cdot 10^3$	0.14908±0.0559	-0.6393±0.0664	0.0
$\hat{b}_2 \cdot 10^3$	-0.016125±0.1233	-0.3379±0.0788	0.1297±0.0124
$\hat{b}_3 \cdot 10^3$	-0.84271±0.1261	-0.7060±0.0719	-0.4166±0.0252
$\hat{b}_4 \cdot 10^3$	1.5127±0.0822	-0.4971±0.0775	0.7942±0.0138
$\hat{c}_1$	1.3882±0.0516	-0.6593±0.0528	Comparison is impossible
$\hat{c}_2$	1.2794±0.0468	-1.0558±0.0622	
$\hat{c}_3$	0.95739±0.0504	0.2150±0.0486	
$\hat{c}_4$	0.39324±0.0330	0.5017±0.0464	
V	$2.16 \cdot 10^{-7}$	$5.13 \cdot 10^{-7}$	$0.92 \cdot 10^{-7}$
$\hat{\sigma}$	$0.658 \cdot 10^{-3}$	$1.013 \cdot 10^{-3}$	$0.428 \cdot 10^{-3}$

Table 5.1 - Parameter estimates from GLS identification of the heat rod.

	Model 1	Model 2	Corresponding model in [14]
Poles	-0.435 0.810+i 0.140 0.810-i 0.140 0.952	-0.189+i 0.672 -0.189-i 0.672 0.806+i 0.269 0.806-i 0.269	0.602+i 0.184 0.602-i 0.184 0.810 0.943
Zeros	-1.166 1.442+i 0.827 1.442-i 0.827	0.064+i 1.089 0.064-i 1.089 -0.652	-1.606+i 1.883 -1.606-i 1.883 -
Static gain	0.2528	-0.0106	0.2440

Table 5.2 - Poles, zeros and static gain of the models of the heat rod process.



<u>Fig. 5.1</u> - Model 1 of the heat-rod process. All variables are given in °C. (Constants are added to the input, the output and the model output.) The sampling period is 10 sec.

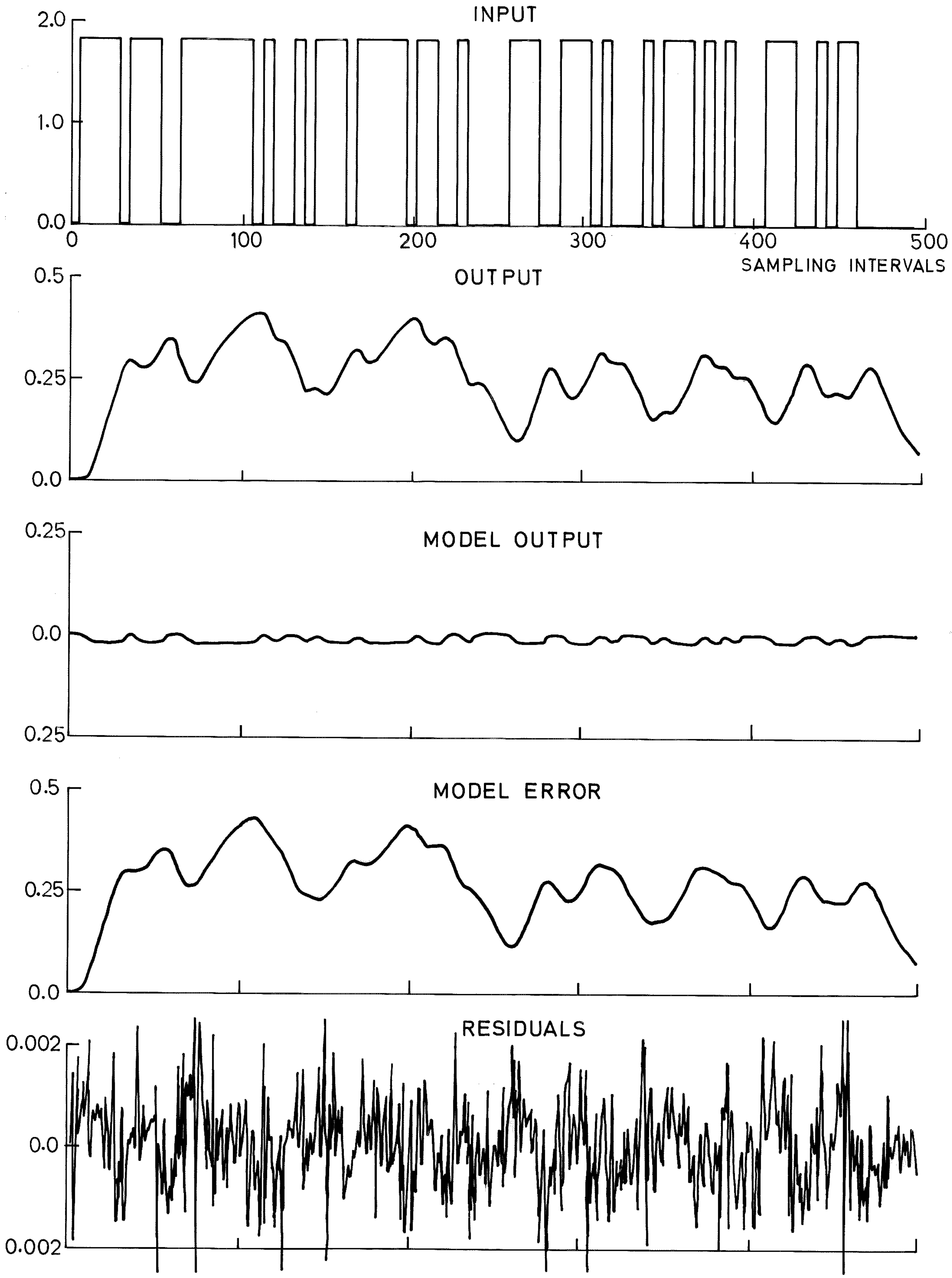


Fig. 5.2 - Model 2 of the heat-rod process. All variables are given in °C. (Constants are added to the input, the output and the model output.) The sampling period is 10 sec.

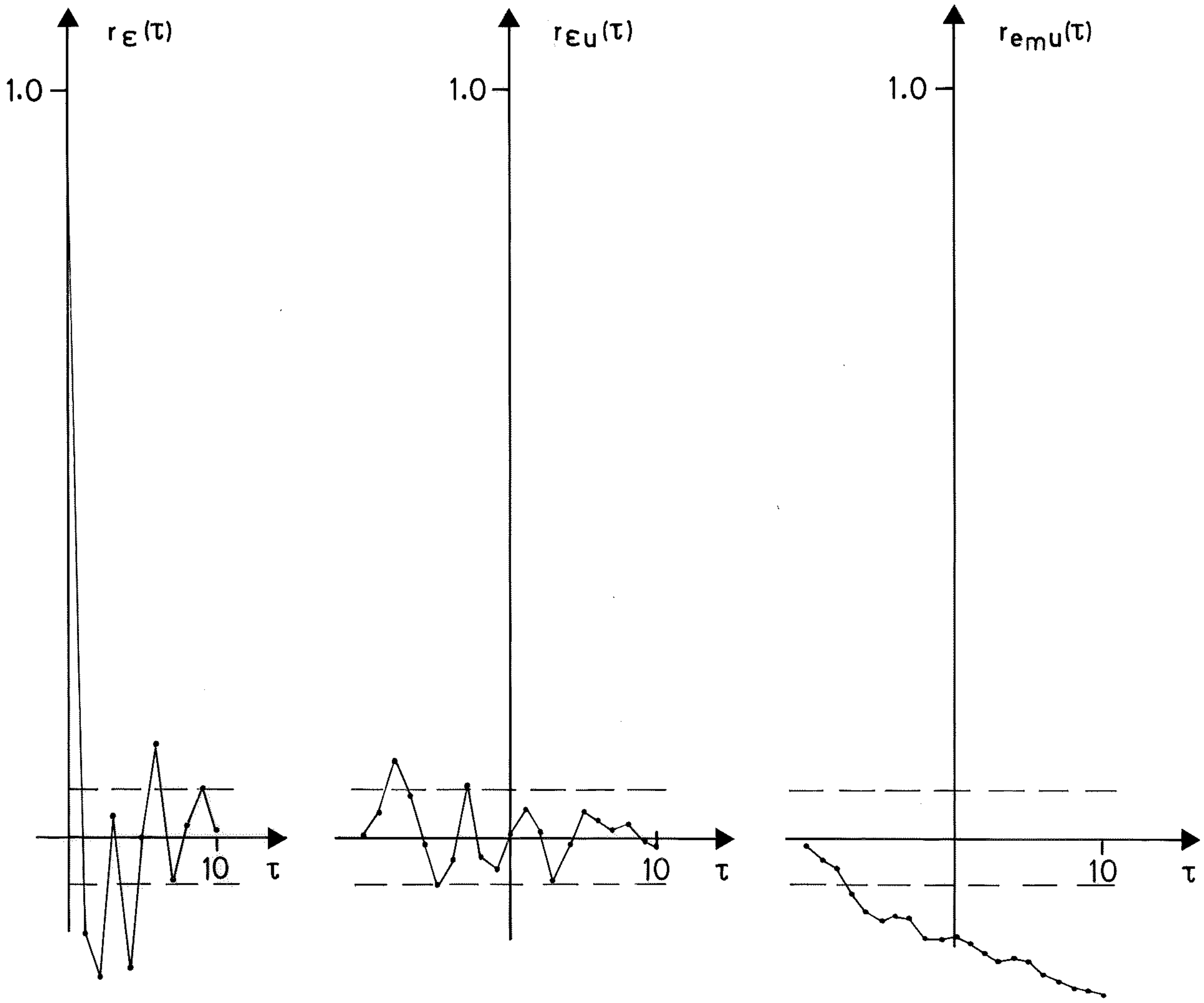


Fig. 5.3 - Normalized sample covariance functions for the heat-rod model 1.
The dashed lines give the 5% confidence interval.
The time is given in sampling periods.

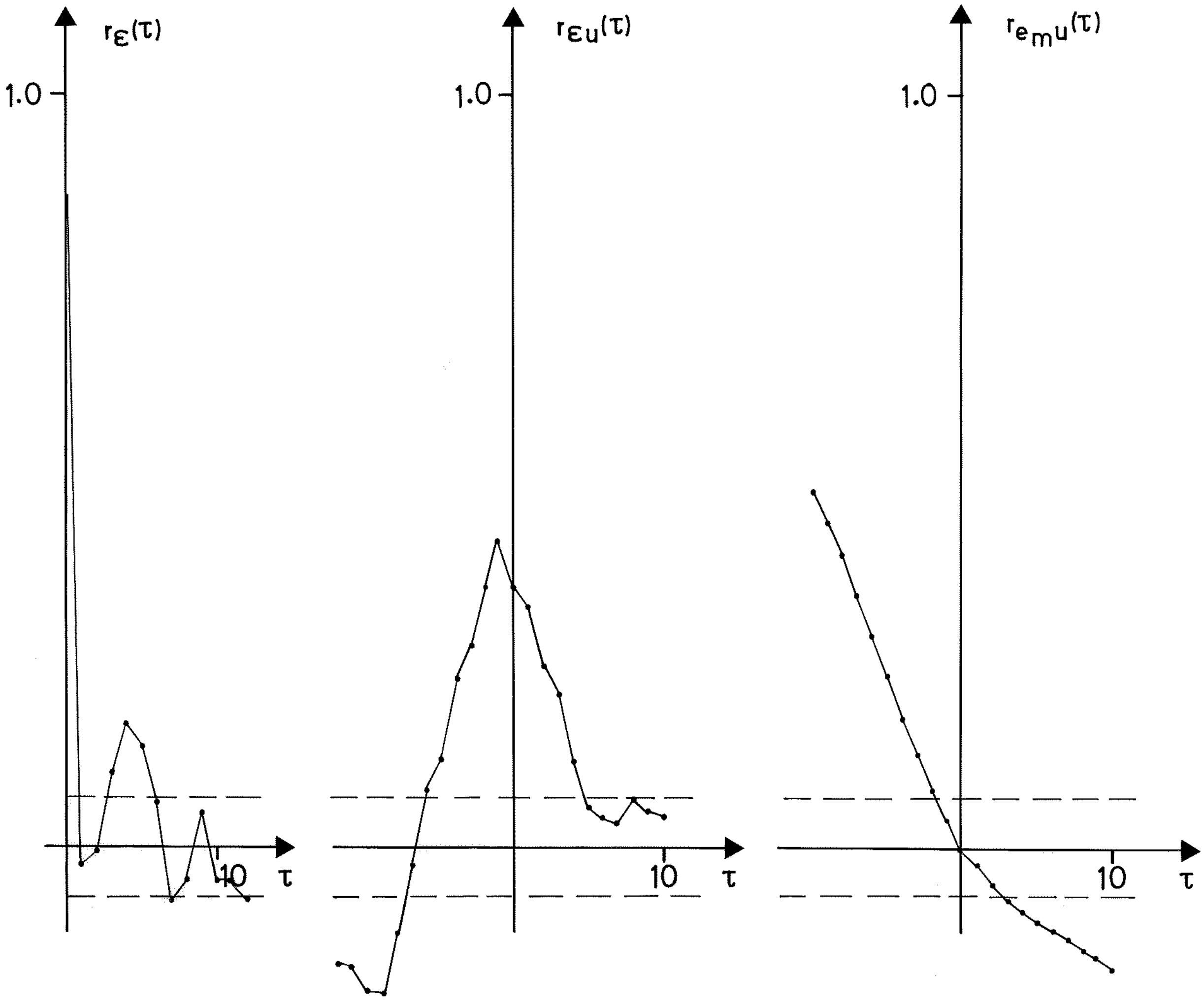


Fig. 5.4 - Normalized sample covariance functions for the heat-rod model 2.
The dashed lines give 5% confidence interval.
The time is given in sampling periods.

5.3. Identification of dynamics of a distillation column.

The system is a binary distillation column. The data have been received from National Physical Laboratory in London. Results of maximum likelihood identifications are reported in [12]. The input signal is the reflux ratio and the output signal is the top product composition. (Experiment 4B, [12], was used.) The test quantity for comparing models of orders 2 and 3, [3], is F(240,3) and has the value 36. Since the ML identification [12] indicates a model of order 2 as reasonable, this order was considered in spite of the great value of the test quantity.

The second order models two minimum points of the loss function were found. The results from the identification are given in Tables 5.3, 5.4 and Figures 5.5 - 5.8.

Discussion of the result.

From Table 5.3 it is seen that $\hat{C}(q^{-1})$ of model 2 is very like $\hat{A}(q^{-1})$ of model 1. With Theorem 3.3 in mind, this is not astonishing.

The model from [12] is very like the model 1, which means that the noise can be well modelled as

$$v(t) = \hat{C}_{ML}(q^{-1})e(t)$$

as well as

$$v(t) = \frac{1}{\hat{C}_{GLS}(q^{-1})} e(t)$$

with e(t) white noise.

The values of the static gain indicate that model 1 gives the best description of the process. Also from the lower value of loss function at the corresponding minimum point, it can be expected that this model is to be preferred.

The plots of the results, Figures 5.5 - 5.6, are a nice illustration of the expected differences.

From Figures 5.7 and 5.8 it is noted that the residuals are most white for model 2 and most uncorrelated with the input for model 1. That means it is hard (or impossible) to choose the "best" model from these figures.

Parameter	Model 1	Model 2	Corresp. ML model in [12]
$\hat{a}_1$	-1.5275±0.0187	0.1865±0.0820	-1.5369±0.0180
$\hat{a}_2$	0.5473±0.0188	-0.0227±0.0529	0.5535±0.0180
$\hat{b}_1$	0.2447±0.0193	0.3730±0.0218	0.2251±0.0190
$\hat{b}_2$	-0.6164±0.0194	0.2174±0.0317	-0.5979±0.0214
$\hat{c}_1$	0.8213±0.062	-1.5076±0.0754	Comparison
$\hat{c}_2$	0.4088±0.059	0.5358±0.0747	impossible
V	69.04	164.37	71.68
$\hat{\sigma}$	11.75	18.13	11.97

Table 5.3 - Parameter estimates from GLS identification of the distillation column data

	Model 1	Model 2	Corresponding model in [12]
Poles	0.95 0.57	-0.093+i 0.118 -0.093-i 0.118	0.96 0.58
Zero	2.52	-0.584	2.66
Static gain	-18.77	0.507	-22.46

Table 5.4 - Poles, zero and static gain of the GLS models of the distillation column.

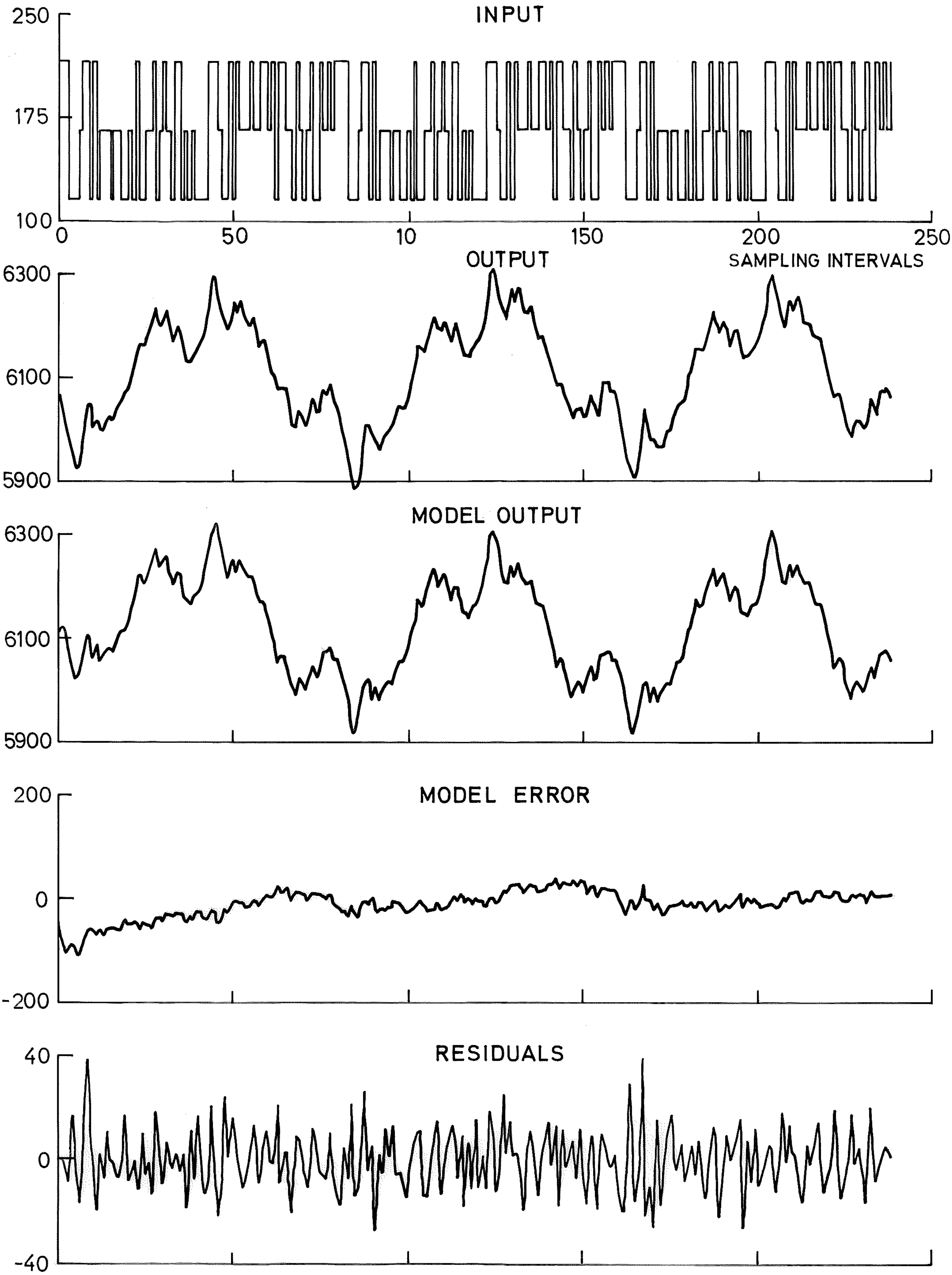


Fig. 5.5 - Model 1 of the distillation column. Digital units are used. The sampling period is 96 sec.

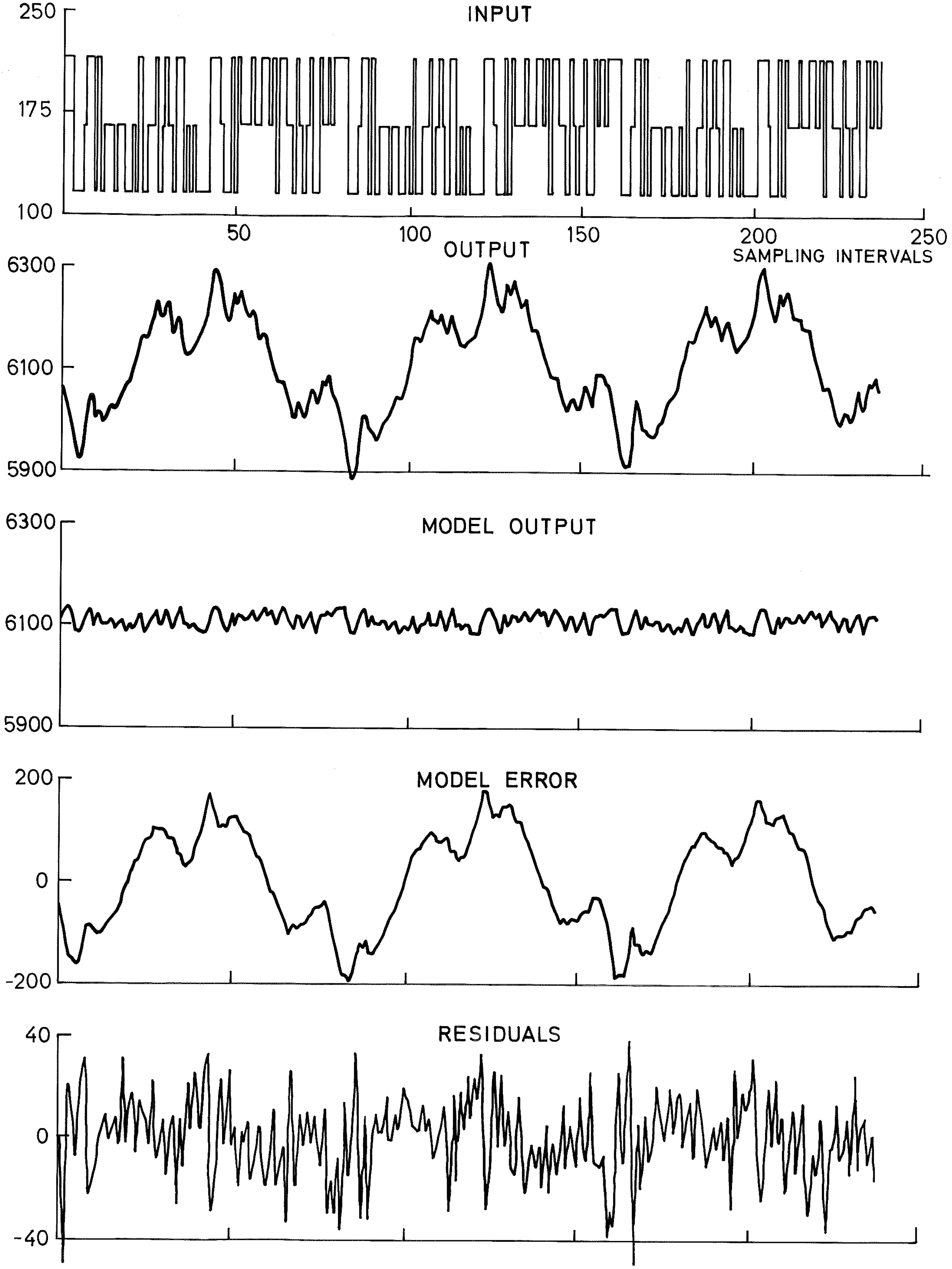


Fig. 5.6 - Model 2 of the distillation column. Digital units are used. The sampling period is 96 sec.

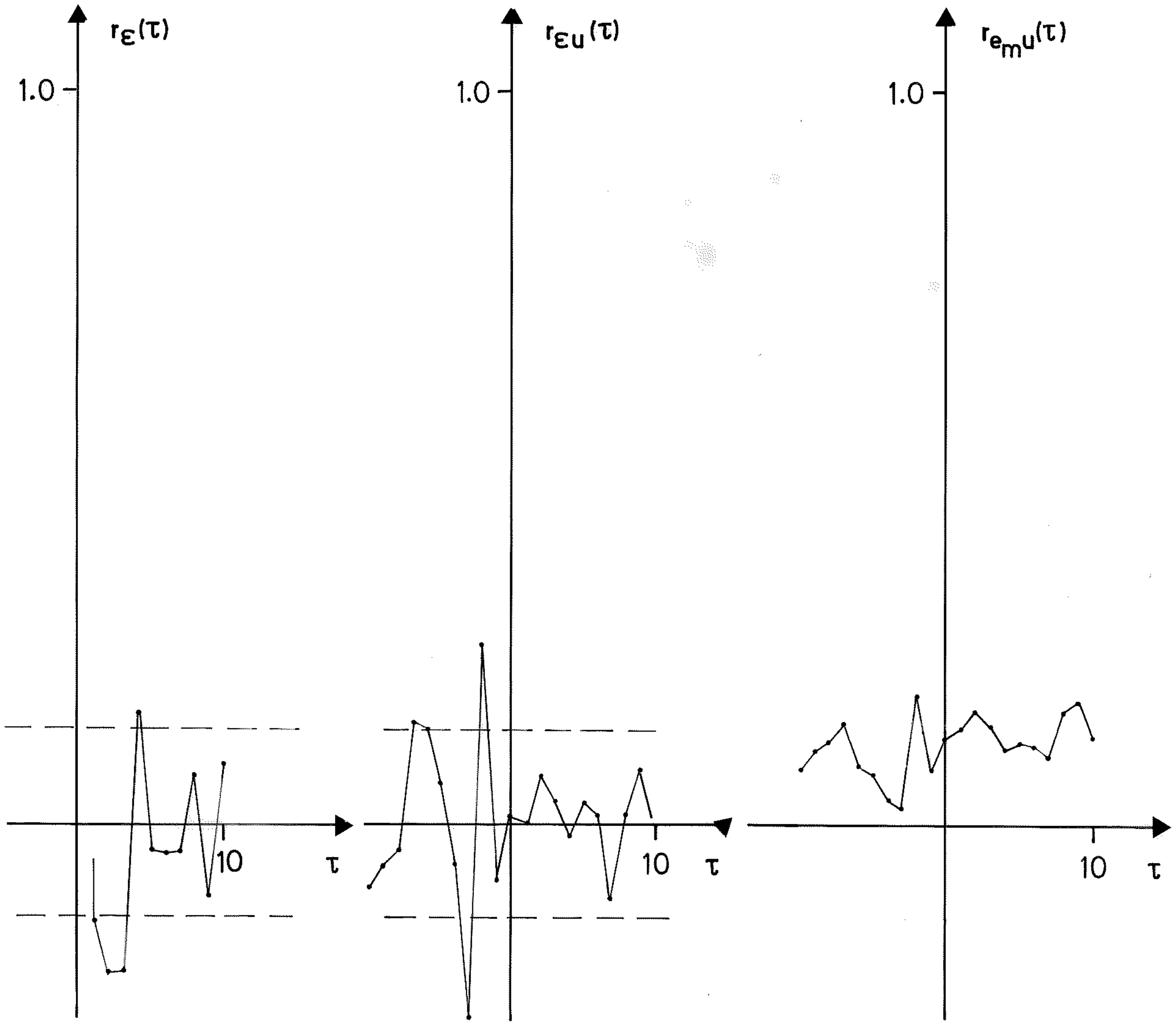


Fig. 5.7 - Normalized sample covariance functions for the distillation column, model 1.
The dashed lines give the 5% confidence interval.
The time is given in sampling periods.

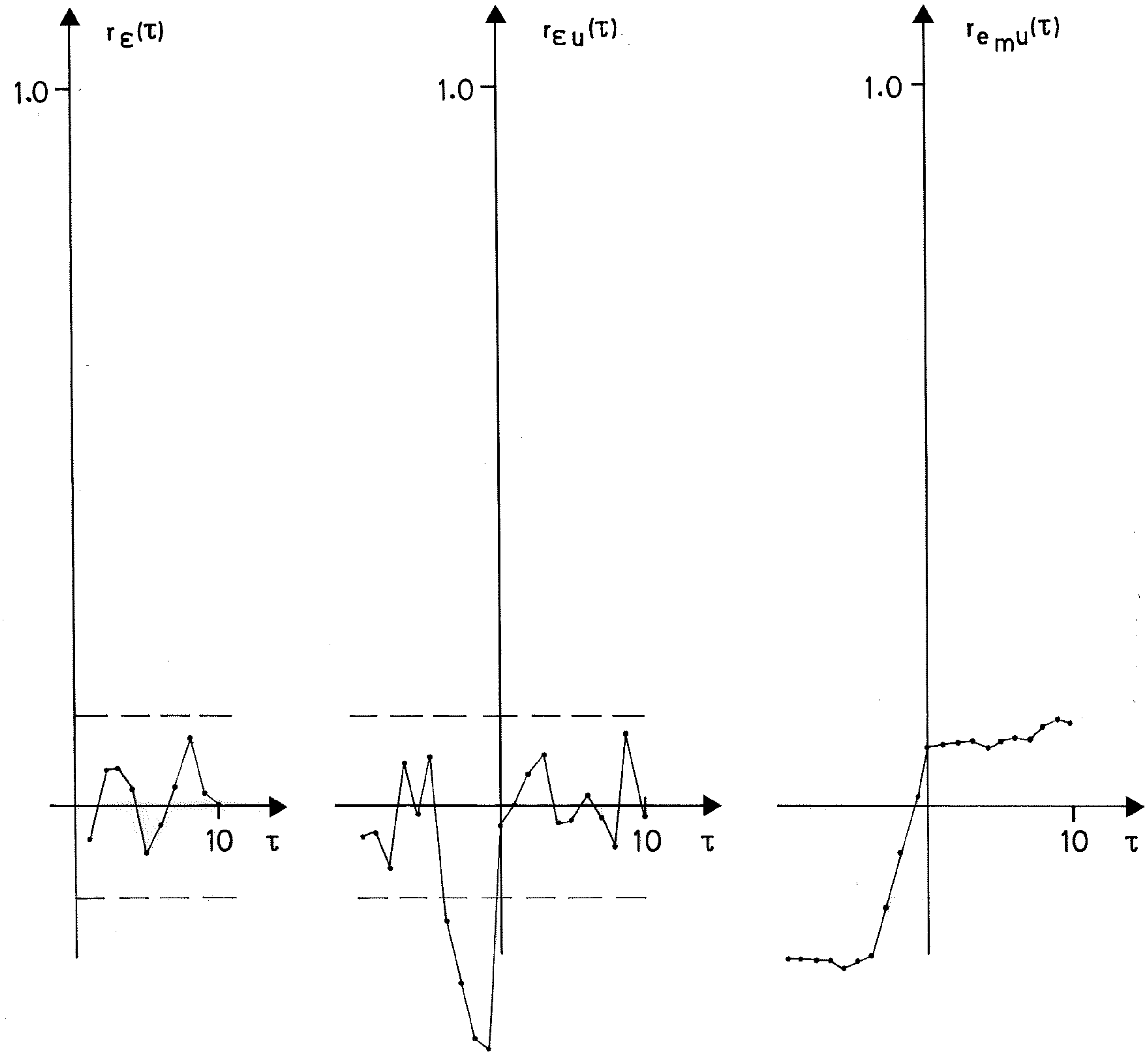


Fig. 5.8 - Normalized sample covariance functions for the distillation column, model 2.
The dashed lines give the 5% confidence interval.
The time is given in sampling periods.

5.4. Identification of dynamics of a nuclear reactor.

The system is a nuclear reactor where the input is reactivity created by control rod movement and the output is the nuclear power, measured by fission chamber. Measurements have been received from OECD Halden Reactor Project in Norway.

The experiment is described in [16] and is called RUN 11 EP 714B. The first 1000 data were used.

The system contains a direct term. This is easily estimated by shifting the input signal. The used $\hat{B}(q^{-1})$ polynomials were of the form

$$\hat{B}(q^{-1}) = \hat{b}_o + \ldots + \hat{b}_n q^{-n}$$

Test quantities for comparing order are F(1000,3). The value when models of orders 1 and 2 are compared is 11.4, while the value is 1.1 when models of orders 2 and 3 are compared. Thus the order two seems to be good.

Two minimum points of the loss function were found for this order. The result of the identifications is given in Tables 5.5, 5.6 and Figures 5.9 - 5.13.

ML identification using the model (5.1) has been done [7], [16].

Discussion of the result.

It is seen from the figures that the differences between the models are small. Further $(\hat{a}_1, \hat{a}_2, \hat{c}_1, \hat{c}_2)$ of model 1 is close to $(\hat{c}_1, \hat{c}_2, \hat{a}_1, \hat{a}_2)$ of model 2. In fact, both models as well as the model in [7] may be simplified to a first order system

$$y(t) = 2.4(1 + 2.6q^{-1})u(t) + \frac{1}{1 - 0.9q^{-1}} e(t) \qquad (5.2)$$

if approximate factors in common and small zeros are omitted.

An identification of a first model gave the result:

$$y(t) = \frac{2.396 + 6.234q^{-1}}{1 - 0.00012q^{-1}} u(t) + \frac{1}{1 - 0.918q^{-1} + 0.0001q^{-2}} e(t)$$

and $\lambda = 2.6660 \cdot 10^{-2}$ which differs just a little from the simplified model.

Since the two models do not differ very much it is impossible to call any of them the "best" or most "correct" one.

If (5.2) is an adequate description of the dynamical behaviour of the process then it is expected with Theorem 3.4 in mind, that there will be (at least) two different but equivalent models of second order. The models obtained by identification are in fact close to these expected models. Of course, this is a very loose discussion according to the assumption that (5.2) describes the system adequately enough.

Parameter	Model 1	Model 2	Model in [7]
$\hat{a}_1$	-0.177±0.062	0.911±0.025	-0.890±0.024
$\hat{a}_2$	0.073±0.019	-0.009±0.021	-0.008±0.019
$\hat{b}_o$	2.41±0.14	2.38±0.14	2.43±0.12
$\hat{b}_1$	5.80±0.21	4.01±0.19	4.02±0.17
$\hat{b}_2$	-1.08±0.38	-5.83±0.16	-5.68±0.15
$\hat{c}_1$	-0.790±0.066	-0.058±0.039	Comparison is
$\hat{c}_2$	-0.124±0.060	0.062±0.034	impossible
$V\cdot 10^4$	3.41	3.47	3.45
$\hat{\sigma}\cdot 10^2$	2.61	2.63	2.63

Table 5.5 - Parameter estimates from identification of the nuclear reactor data.

	Model 1	Model 2	Model in [7]
Poles	0.088+i 0.256 0.088-i 0.256	-0.010 0.921	-0.009 0.898
Zeros	-2.575 0.174	-2.621 0.935	-2.564 0.911
Static gain	7.95	7.07	7.50

Table 5.6 - Poles, zeros and static gain of the models of the nuclear reactor.

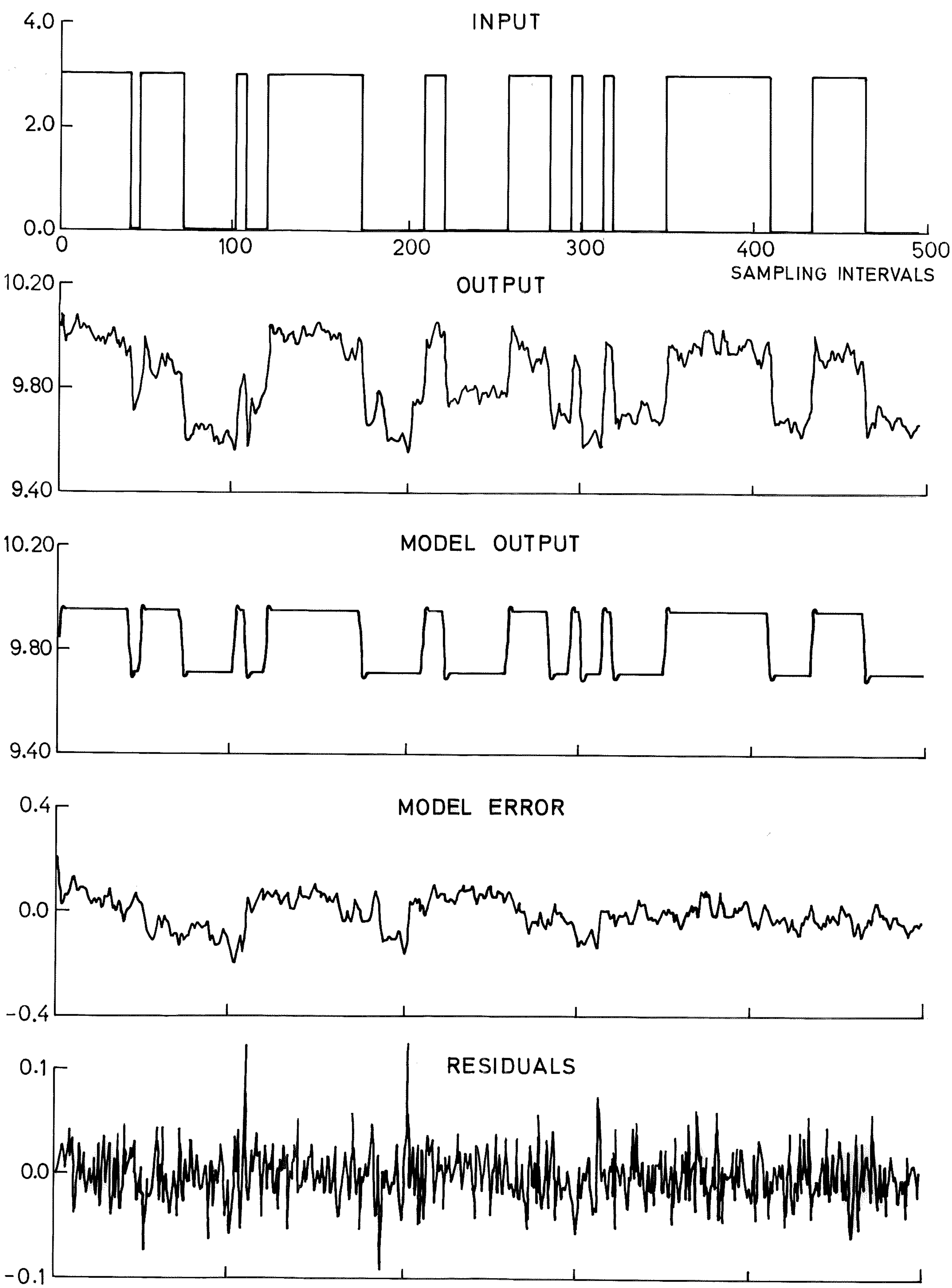


Fig. 5.9 - Model 1 of the nuclear reactor. The input is given in digital units and the other variables in MW. The sampling period is 2 seconds.

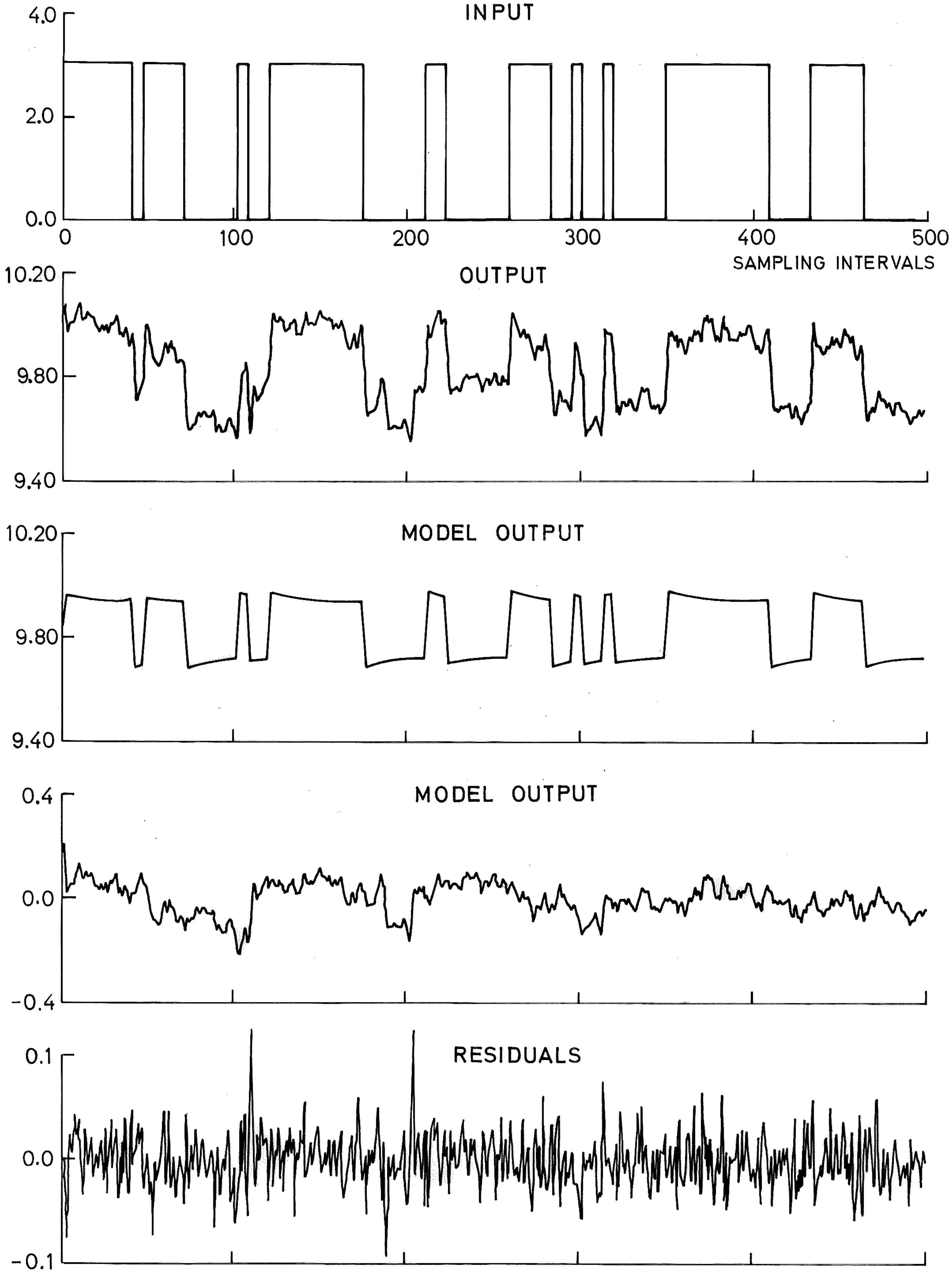


Fig. 5.10 - Model 2 of the nuclear reactor. The input is given in digital units and the other variables in MW. The sampling period is 2 seconds.

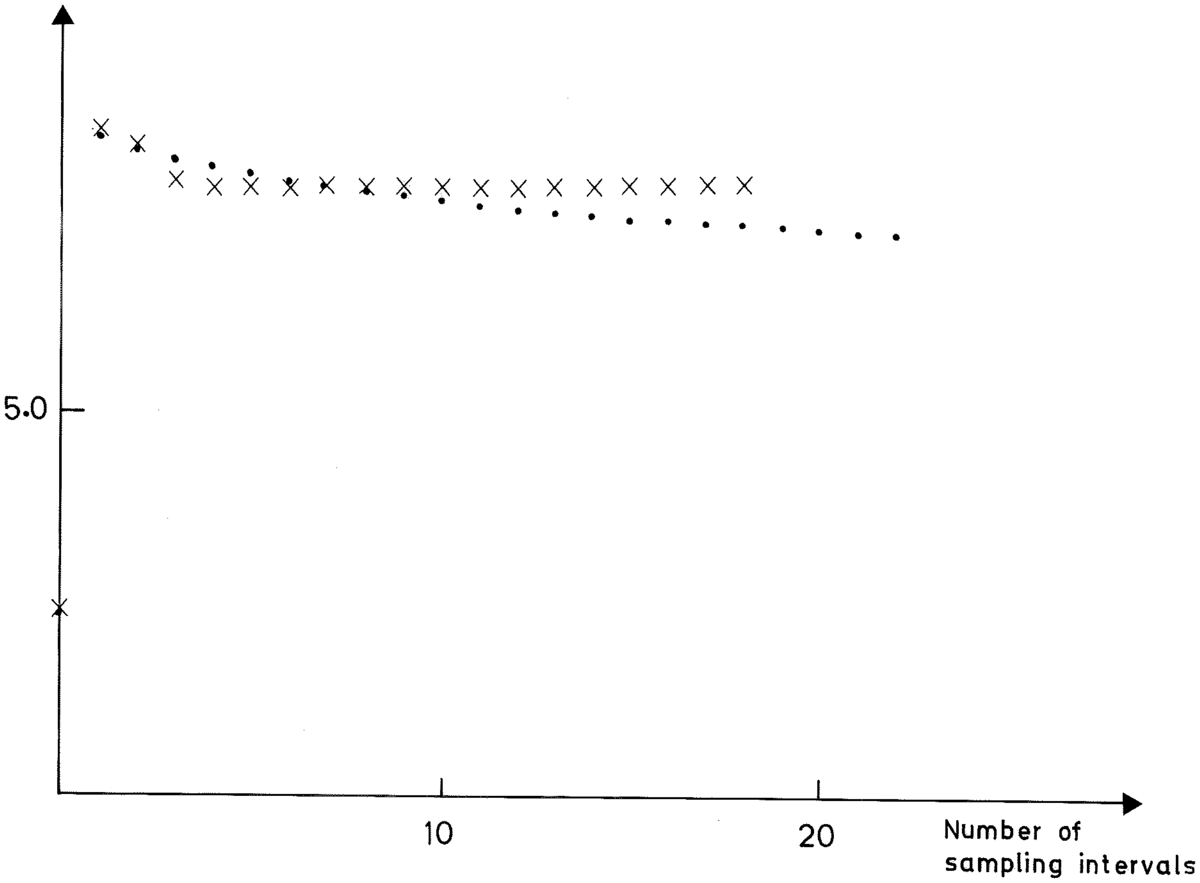


Fig. 5.11 - Step response of the nuclear reactor models.
x = model 1
· = model 2

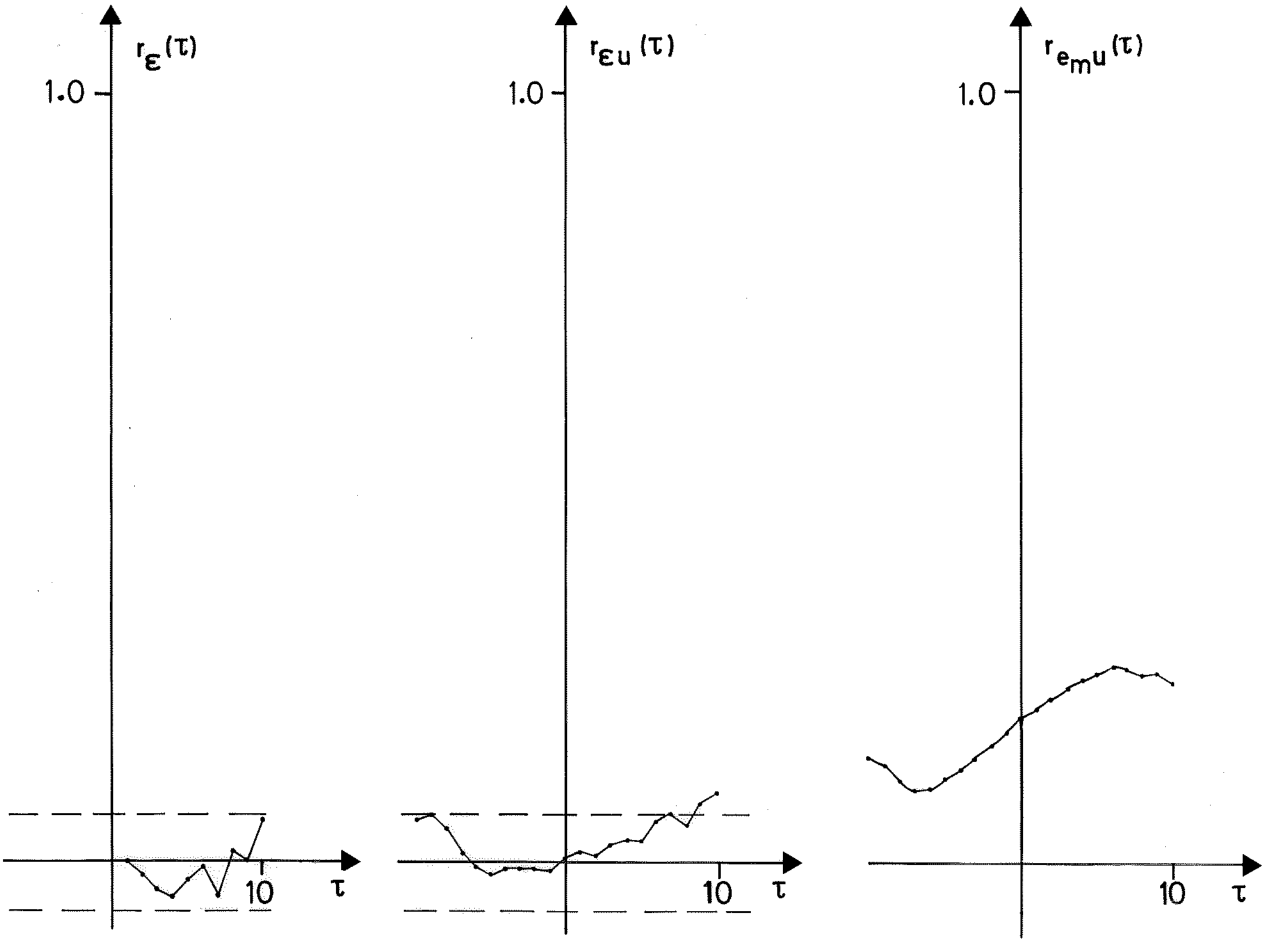


Fig. 5.12 - Normalized sample covariance functions for the nuclear reactor, model 1.
The dashed lines give the 5% confidence interval.
The time is given in sampling periods.

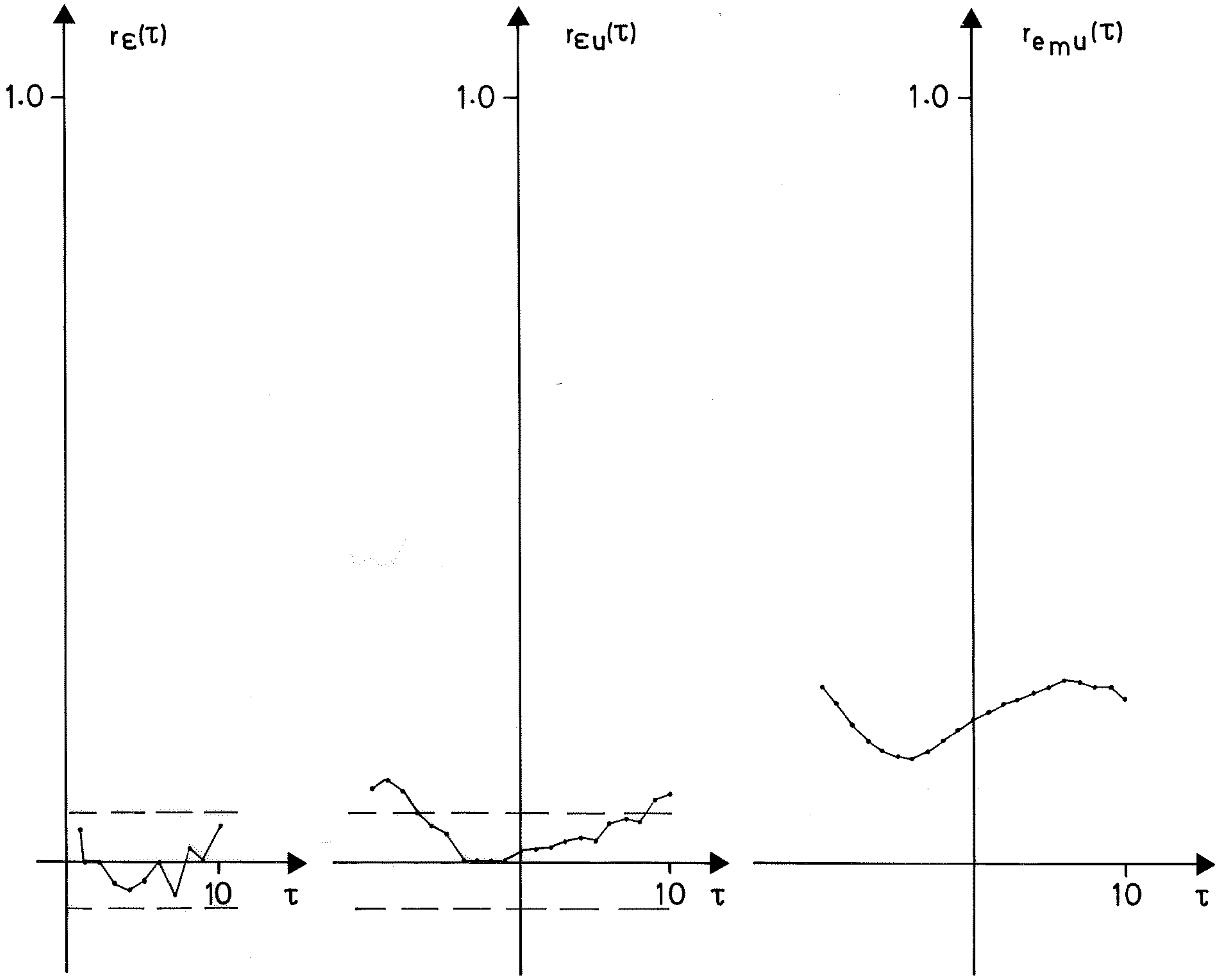


Fig. 5.13 - Normalized sample covariance functions for the nuclear reactor, model 2.
The dashed lines give the 5% confidence interval.
The time is given in sampling periods.

VI. CONCLUSIONS.

Some essential properties of the generalized least squares (GLS) method for identification of dynamical systems are summarized below. Part of the material is well-known.

o The GLS method can be interpreted as a maximum likelihood method when suitable assumptions of the structure of the equations governing the system are made. The estimation of the correlation of the noise can be done in some different ways, and every way to do it corresponds to some structure of the system equations. The GLS method is then a special minimization algorithm applied to a corresponding loss function, which is a sum of squared residuals. The GLS method will then have nice asymptotic properties.

o The GLS method is an uncomplicated extension of the least squares (LS) method. Besides a program for LS identification only programs for administration and filtering are needed.

o The GLS method gives a very slow convergence close to a minimum point of the loss function. The method is thus inappropriate if great accuracy is required.

o Applied to nice data the GLS method gives good results comparable with the results of the more complicated "ordinary" maximum likelihood method. The required conditions of the data are weaker than for the simpler LS method. For a sufficiently high value of the signal to noise ratio it can be shown theoretically that the loss function has only one local minimum point.

o The loss function corresponding to the GLS method may have more than one minimum point. In this case the result of the GLS identification depends on the start values of the parameter estimates. The existence of several minimum points can be shown theoretically for low signal to noise ratios. In practice it can happen also for reasonable values of this ratio. It is not always easy without intimate knowledge of the actual process to decide which of the models that will be the "best" or most "correct".

VII. ACKNOWLEDGEMENTS.

The author wants to express his great gratitude to Prof. K.J. Åström for suggesting the subject and for his valuable guidance.

He also wants to thank his colleagues, tekn.lic. Ivar Gustavsson and civ.ing. Lennart Ljung, for many stimulating discussions.

It is a pleasure to thank Mrs. G. Christensen, who typed the manuscript, and Mrs. B. Tell, who drew the figures.

The author is grateful for the measurements, which were supplied by tekn.lic. Bo Leden and National Laboratory, London, and OECD Halden Reactor Project.

VIII. REFERENCES.

[1] K.J. Åström: Lectures on the Identification Problem - the Least Squares Method. Report 6806, 1968, Division of Automatic Control, Lund Institute of Technology.

[2] K.J. Åström: Introduction to Stochastic Control Theory. Academic Press, 1970.

[3] K.J. Åström and T. Bohlin: Numerical Identification of Linear Dynamic Systems from Normal Operating Records. Paper, IFAC Symposium on Theory of Self-Adaptive Systems, Teddington, England. In Theory of Self-Adaptive Control Systems (Ed. P.H. Hammond), Plenum Press, New York, 1966.

[4] K.J. Åström and P. Eykhoff: System Identification - A Survey. Automatica 7, 123 - 162, 1971.

[5] T. Bohlin: On the Problem of Ambiguites in Maximum Likelihood Identification. Automatica 7, 199 - - 210, 1971.

[6] P.E. Caines: The Parameter Estimation of State Variable Models of Multivariable Linear Systems. Control Systems Centre Report No. 146, The University of Manchester, Institute of Science and Technology, 1971.

[7] S. Carlsson: Maximum Likelihood identifiering av reaktordynamik från flervariabla experiment. Master Thesis, Division of Automatic Control, Lund Institute of Technology.

[8] D.W. Clarke: Generalized Least Squares Estimation of the Parameters of a Dynamic Model. 1st IFAC Symposium on Identification in Automatic Control Systems, Prague, 1967.

[9] I.H. Cramér and M.R. Leadbetter: Stationary and Related Stochastic Processes. John Wiley & Sons, New York, 1967.

[10] D.K. Faddeev and V.N. Faddeeva: Computational Methods of Linear Algebra. W.H. Freeman and Company, San Fransisco, 1963.

[11] B.V. Gnedenko: The Theory of Probability. Chelsea Publishing Company, New York, 1963.

[12] I. Gustavsson: Identification of Dynamics of a Distillation Column. Report 6916, 1969, Division of Automatic Control, Lund Institute of Technology.

[13] E. Kreindler and A. Jameson: Conditions for Nonnegativeness of Partioned Matrices. IEEE Trans. Aut. Control, Feb., 1972, 147 - 148.

[14] B. Leden: Identification of Dynamics of a One Dimensional Heat Diffusion Process. Report 7121, 1971, Division of Automatic Control, Lund Institute of Technology.

[15] L. Ljung: Characterization of the Concept of "Persistently Exciting" in the Frequency Domain. Report 7119, 1971, Division of Automatic Control, Lund Institute of Technology.

[16] G. Olsson: Identification of the Halden Boiling Water Reactor Dynamics. Report, Division of Automatic Control, Lund Institute of Technology. To appear.

[17] J.M. Ortega, W.C. Rheinboldt: Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, New York, 1970.

[18] P.H. Phillipson: Convergence of Clarke's Generalized Least Squares Method in Process Parameter Estimation. 1970, Dep. of Eng., Univ. of Leicester, England.

[19] R. Rao and S.K. Mitra: Generalized Inverse of Matrices and its Applications. John Wiley & Sons, New York, 1971.

[20] N. Wiener: Generalized Harmonic Analysis and Tauberian Theorems. The MIT Press, MIT, Mass., 1964.

APPENDIX A

A SUMMARY OF ERGODICITY THEOREMS.

The purpose of this appendix is a study of expressions of the type

$$\frac{1}{n}\sum_{t=1}^{n} z_1(t)z_2(t)$$

and their limits as $n \to \infty$. $z_i(t)$ will be deterministic signals or stationary stochastic processes of the type

$$z(t) = H(q^{-1})e(t)$$

where $H(q^{-1})$ is a stable filter and $e(t)$ a sequence of independent, equally distributed random variables (white noise). For the study of such expressions some well-known ergodicity theorems will be used. In order to show how these are exploited, the theorems will be stated here in form of two lemmas.

Lemma A.1: Assume that $x(t)$ is a stationary process with discrete time and finite variance. If the covariance function $r_x(\tau) \to 0$ as $|\tau| \to \infty$ then

$$\frac{1}{n}\sum_{t=1}^{n} x(t) \to Ex(t)$$

with probability one and in mean square.

Proof: See [11].

Lemma A.2: Assume that $x(t)$ is a stochastic process with zero mean. If the covariance function fulfils

$$|r(t,s)| \leqslant K \frac{t^{\alpha} + s^{\alpha}}{1 + |t-s|^{\beta}}$$

with $K > 0$, $0 \leqslant 2\alpha < \beta < 1$ then

$$\frac{1}{n} \sum_{t=1}^{n} x(t) \to 0$$

with probability one and in mean square.

Proof: See [9].

Some kinds of conditions for deterministic signals are also needed. Inspired of the theory of almost periodic functions, see [20], almost periodic sequences will be used. In the time discrete case the results are much more simple than for time continuous functions.

Definition A.1: The sequence $\{u(t)\}_{t=1}^{\infty}$ is said to be almost periodic if to every $\varepsilon > 0$ there exists a periodic sequence $\{v(t)\}_{t=1}^{\infty}$ (that is $v(t) = v(t+T)$ some T, all t) with finite period T, such that

$$|v(t) - u(t)| < \varepsilon \text{ all } t$$

It is now possible to start the analysis.

Lemma A.3: Let the stationary stochastic processes $z_1(t)$ and $z_2(t)$ be given by

$$z_1(t) = G(q^{-1})e(t)$$

$$z_2(t) = H(q^{-1})e(t)$$

where $e(t)$ is white noise with zero mean, unit variance and finite fourth moment μ.

$$G(q^{-1}) = \sum_{i=0}^{\infty} g_i q^{-i}$$

and

$$H(q^{-1}) = \sum_{i=0}^{\infty} h_i q^{-i}$$

If

$$\sum_{i=0}^{\infty} g_i^2 < \infty, \quad \sum_{i=0}^{\infty} h_i^2 < \infty$$

then

$$\frac{1}{n} \sum_{t=1}^{n} z_1(t)z_2(t) \rightarrow Ez_1(t)z_2(t) = \sum_{i=0}^{\infty} h_i g_i, \qquad n \rightarrow \infty$$

with probability one and in mean square.

Remark: The condition on $G(q^{-1})$ and $H(q^{-1})$ means just that $z_1(t)$ and $z_2(t)$ have finite variances.

Proof: Define $v(t) = z_1(t) \cdot z_2(t)$.

$v(t)$ is a stationary stochastic process with

$$Ev(t) = \sum_{i=0}^{\infty} h_i g_i$$

The convergence of this sum is an immediate consequence of the assumptions and Schwartz' lemma.

In order to use Lemma A.1 the covariance function must be computed.

$$r_v(\tau) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} g_i g_j h_k h_\ell Ee(t-i)e(t+\tau-j) \cdot$$

$$\cdot\ e(t-k)e(t+\tau-\ell) - \left(\sum_{i=0}^{\infty} h_i g_i\right)^2$$

But

$$Ee(t-i)e(t+\tau-j)e(t-k)e(t+\tau-\ell) =$$

$$= \delta_{j,\tau+i}\delta_{\ell,\tau+k} + \delta_{i,k}\delta_{j,\ell} + \delta_{\ell,\tau+i}\delta_{j,\tau+k} +$$

$$+ (\mu-3)\delta_{j,\tau+i}\delta_{k,i}\delta_{\ell,\tau+i}$$

which gives

$$r_v(\tau) = \sum_{i=0}^{\infty} \sum_{k=0}^{\infty} g_i g_{i+\tau} h_k h_{k+\tau} +$$

$$+ \sum_{i=0}^{\infty} \sum_{k=0}^{\infty} g_i g_{k+\tau} h_k h_{i+\tau} + (\mu-3) \sum_{i=0}^{\infty} g_i g_{i+\tau} h_i h_{i+\tau} =$$

$$= \left(\sum_{i=0}^{\infty} g_i g_{i+\tau}\right)\left(\sum_{k=0}^{\infty} h_k h_{k+\tau}\right) +$$

$$+ \left(\sum_{i=0}^{\infty} g_i h_{i+\tau}\right)\left(\sum_{k=0}^{\infty} g_{k+\tau} h_k\right) +$$

$$+ (\mu-3) \sum_{i=0}^{\infty} g_i g_{i+\tau} h_i h_{i+\tau}$$

From this expression the following inequalities are obtained using Schwartz' lemma.

$$|r_v(\tau)| \leq \sqrt{\sum_{i=0}^{\infty} g_i^2 \sum_{j=0}^{\infty} g_{i+\tau}^2 \sum_{k=0}^{\infty} h_k^2 \sum_{\ell=0}^{\infty} h_{\ell+\tau}^2} \; +$$

$$+ \sqrt{\sum_{i=0}^{\infty} g_i^2 \sum_{j=0}^{\infty} h_{j+\tau}^2 \sum_{k=0}^{\infty} g_{k+\tau}^2 \sum_{\ell=0}^{\infty} h_{\ell}^2} \; +$$

$$+ |\mu-3| \sqrt{\sum_{i=0}^{\infty} g_i^2 h_i^2 \sum_{j=0}^{\infty} g_{j+\tau}^2 h_{j+\tau}^2}$$

But

$$\sum_{i=0}^{\infty} g_{i+\tau}^2 = \sum_{i=0}^{\infty} g_i^2 - \sum_{i=0}^{\tau-1} g_i^2 \to 0$$

as $\tau \to \infty$, which implies $|r_v(\tau)| \to 0$, as $\tau \to \infty$.

Invoking Lemma A.1 the proof is finished.

Q.E.D.

Lemma A.4: Let the stationary stochastic processes $z_1(t)$ and $z_2(t)$ be given by

$$z_1(t) = G(q^{-1}) \cdot e_1(t)$$

$$z_2(t) = H(q^{-1}) \cdot e_2(t)$$

Here $e_1(t)$ and $e_2(t)$ are independent white noises with zero means and unit variances.

$$G(q^{-1}) = \sum_{i=0}^{\infty} g_i q^{-i}$$

and

$$H(q^{-1}) = \sum_{i=0}^{\infty} h_i q^{-i}$$

If

$$\sum_{i=0}^{\infty} g_i^2 < \infty, \quad \sum_{i=0}^{\infty} h_i^2 < \infty$$

then

$$\frac{1}{n} \sum_{t=1}^{n} z_1(t) z_2(t) \to 0, \quad n \to \infty$$

with probability one and in mean square.

Proof: Define $v(t) = z_1(t) \cdot z_2(t)$, a stochastic process with zero mean.

The covariance function of $v(t)$ is

$$r_v(\tau) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} g_i g_j h_k h_\ell E e_1(t-i) e_1(t+\tau-j) \cdot$$

$$\cdot\ e_2(t-k) e_2(t+\tau-\ell) =$$

$$= \sum_{i=0}^{\infty} g_i g_{i+\tau} \sum_{k=0}^{\infty} h_k h_{k+\tau}$$

From this expression

$$|r_v(\tau)| \leq \sqrt{\sum_{i=0}^{\infty} g_i^2 \sum_{j=0}^{\infty} g_{j+\tau}^2 \sum_{k=0}^{\infty} h_k^2 \sum_{\ell=0}^{\infty} h_{\ell+\tau}^2} \to 0, \quad \tau \to \infty$$

The assertion of the lemma now follows from Lemma A.1.

Q.E.D.

Lemma A.5: Let $z_1(t)$ be a deterministic, bounded sequence and $z_2(t)$ a stationary, stochastic process, given by

$$z_2(t) = G(q^{-1})e(t)$$

where $e(t)$ is white noise with zero mean and unit variance,

$$G(q^{-1}) = \sum_{i=0}^{\infty} g_i q^{-i}, \qquad \sum_{i=0}^{\infty} g_i^2 < \infty$$

If the covariance function of $z_2(t)$ fulfils

$$|r_{z_2}(\tau)| \leqslant C\tau^{-\gamma} \qquad \gamma > 0 \qquad \tau \geqslant 1$$

then

$$\frac{1}{n}\sum_{t=1}^{n} z_1(t)z_2(t) \to 0, \qquad n \to \infty$$

with probability one and in mean square.

Proof: Define $v(t) = z_1(t) \cdot z_2(t)$, a stochastic process with zero mean. By the assumptions $z_1(t)$ is bounded, say $|z_1(t)| \leqslant D$. The covariance function of $v(t)$ fulfils

$$|r_v(t,s)| = |Ez_1(t)z_2(t)z_1(s)z_1(s)| \leqslant D^2|r_{z_2}(t-s)| \leqslant$$

$$\leqslant D^2C|t-s|^{-\gamma} \qquad \text{for } |t-s| \geqslant 1$$

Lemma A.2 can now be used with $\alpha = 0$, $\beta = \gamma$ and

$$K = D^2\max\left(\frac{r_{z_2}(0)}{2}, C\right)$$

Q.E.D.

Lemma A.6: Let $z_1(t)$ and $z_2(t)$ be two almost periodic sequences. Then

$$\frac{1}{n} \sum_{t=1}^{n} z_1(t) \cdot z_2(t)$$

converges as $n \to \infty$.

Proof: Define $v(t) = z_1(t) \cdot z_2(t)$. Clearly $v(t)$ is also almost periodic. Let $u(t)$ be a periodic sequence such that

$$|v(t) - u(t)| < \varepsilon \quad (\text{all } t)$$

The convergence of

$$\frac{1}{n} \sum_{t=1}^{n} u(t)$$

is trivial. Put

$$s_n = \frac{1}{n} \sum_{t=1}^{n} v(t)$$

Using the Cauchy criterion for the sequence

$$\frac{1}{n} \sum_{t=1}^{n} u(t)$$

$$|s_n - s_m| = \left| \frac{1}{n} \sum_{t=1}^{n} (v(t) - u(t) + u(t)) - \right.$$

$$\left. - \frac{1}{m} \sum_{t=1}^{m} (v(t) - u(t) + u(t)) \right| \leqslant$$

$$\leqslant 2\varepsilon + \left| \frac{1}{n} \sum_{t=1}^{n} u(t) - \frac{1}{m} \sum_{t=1}^{m} u(t) \right| < 3\varepsilon$$

if $\min(m,n) > N(\varepsilon)$

Using the same criterion for the sequence s_n the convergence is proved.

Q.E.D.

The following example shows that x(t) bounded does not imply convergence of

$$\frac{1}{n} \sum_{t=1}^{n} x(t)$$

This means especially that $z_i(t)$ bounded is a too weak condition in Lemma A.6.

Example: Define x(t) by

$$\begin{aligned} x(t) &= 1 & t &= 1 \\ &= -1 & t &= 2,.4 \\ &= 1 & t &= 3...12 \\ &= -1 & t &= 13...36 \end{aligned}$$

and

$$x(t) = (-1)^{m-1}, \quad 4 \cdot 3^{m-1} + 1 \leqslant t \leqslant 4 \cdot 3^{m}$$

Put

$$s_n = \frac{1}{n} \sum_{t=1}^{n} x(t)$$

Then $s_n = 1/2$ if $n = 4 \cdot 3^m$ m odd

and $s_n = -1/2$ if $n = 4 \cdot 3^m$ m even

From this it follows that lim inf s_n < lim sup s_n and thus lim s_n does not exist.

It is now possible to prove Theorem 2.1.

Proof of Theorem 2.1: An inspection of the kind of terms in (2.1) shows that the proof follows immediately from Lemmas A.3 - A.6.

If e(t) and/or v(t) has not zero mean, it is rewritten as e(t) = [e(t) - Ee(t)] + Ee(t) and the lemmas are applied twice. In this case the following easily proved property is required as well.

If $v(t) = H(q^{-1}) \cdot e(t)$, e(t) white noise with zero mean and

$$\sum_{i=0}^{\infty} h_i^2 < \infty$$

then

$$\frac{1}{n} \sum_{t=1}^{n} v(t) \to 0, \quad n \to \infty$$

with probability one and in mean square.

Q.E.D.

APPENDIX B
ANALYSIS OF THE MINIMIZATION ALGORITHM.

The purpose of this appendix is to examine the properties of the minimization algorithm. To get reasonable work it is assumed that the loss function is a quadratic form, which is a good approximation close to a stationary point.

Define

$$W(x,y) = \frac{1}{2}[x^T \quad y^T]\begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} \tag{B.1}$$

where Q is a symmetric matrix.

The vector x corresponds to $[\hat{a}_1-a_1, \ldots, \hat{b}_n-b_n]^T$ and the vector y corresponds to $[\hat{c}_1-c_1, \ldots, \hat{c}_n-c_n]^T$.

The minimization procedure is given by

$$\begin{cases} Q_{11}x_{k+1} + Q_{12}y_k = 0 \\ \\ Q_{21}x_{k+1} + Q_{22}y_{k+1} = 0 \end{cases} \tag{B.2}$$

It is assumed in the following that $Q_{11} > 0$, $Q_{22} > 0$ (are positive definite) which always can be assumed to be true for the actual loss function (3.3). An exception is the case of no noise and too high an order of the model, but this case can be excluded. This means that (B.2) has always a unique solution.

Introduce

$$\begin{cases} P_1 = Q_{11}^{-1}Q_{12}Q_{22}^{-1}Q_{21} \\ P_2 = Q_{22}^{-1}Q_{21}Q_{11}^{-1}Q_{12} \end{cases} \tag{B.3}$$

Then from (B.2)

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix} \begin{bmatrix} x_k \\ y_k \end{bmatrix} \tag{B.4}$$

It is of great interest to analyze the eigenvalues of the matrix

$$\begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}$$

Lemma B.1: Let A and B be two matrices, such that AB and BA are defined. If $\lambda \neq 0$ is an eigenvalue of AB then λ is also an eigenvalue of BA.

Proof: $ABe = \lambda e$ gives $BABe = \lambda Be$.
If $Be \neq 0$ then λ is an eigenvalue of BA with the eigenvector Be.
If $Be = 0$ then $\lambda = 0$, a contradiction.

Q.E.D.

Corr: P_1 and P_2 have the same non-zero eigenvalues.

The following well-known lemma will be used below and in Appendix C.

Lemma B.2: The symmetric matrix

$$Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$$

is positive definite if and only if $Q_{22} > 0$ and $Q_{11} - Q_{12}Q_{22}^{-1}Q_{21} > 0$.

Further, if $Q \geqslant 0$ (positive semidefinite) and $Q_{22} > 0$ then $Q_{11} - Q_{12}Q_{22}^{-1}Q_{21}$ is positive semidefinite.

Proof: See [13].

Introduce

$$\begin{cases} \tilde{Q}_1 = Q_{11} - Q_{12}Q_{22}^{-1}Q_{21} \\ \tilde{Q}_2 = Q_{22} - Q_{21}Q_{11}^{-1}Q_{12} \end{cases} \tag{B.5}$$

Then the criterion (with the assumptions above of Q_{11} and Q_{22}) can be written

$Q > 0$ if and only if $\tilde{Q}_1 > 0$ if and only if $\tilde{Q}_2 > 0$

(B.3) and (B.5) give easily

$$\begin{cases} P_1 = I - Q_{11}^{-1}\tilde{Q}_1 \\ P_2 = I - Q_{22}^{-1}\tilde{Q}_2 \end{cases} \tag{B.6}$$

Let P_1 have an eigenvalue λ with the associated eigenvector e

$$P_1 e = \lambda e$$

(B.6) gives

$$e - Q_{11}^{-1}\tilde{Q}_1 e = \lambda e, \quad \tilde{Q}_1 e = (1 - \lambda)Q_{11}e$$

and

$$1 - \lambda = \frac{e^T\tilde{Q}_1 e}{e^T Q_{11} e} \tag{B.7}$$

Lemma B.3: All eigenvalues of P_1 are positive.

Proof:

$$e^T\tilde{Q}_1 e = e^T Q_{11} e - e^T Q_{12} Q_{22}^{-1} Q_{21} e \leqslant e^T Q_{11} e$$

which gives

$$1 - \lambda \leqslant 1 \quad \text{or} \quad \lambda \geqslant 0$$

Q.E.D.

Lemma B.4: P_1 has a basis of eigenvectors.

Proof: Follows from [19] (Thm. 6.2.3) since P_1 is a product of a positive definite matrix and a positive (semi-) definite matrix.

Lemma B.5: Let λ denote an eigenvalue of P_1.

i) $Q > 0$ if and only if $\lambda < 1$ (all λ).

ii) $Q \geqslant 0$ if and only if $\lambda \leqslant 1$ (all λ) with equality for at least one λ.

iii) Q indefinite if and only if $\lambda > 1$ some λ.

Proof:

i) Assume $Q > 0$. Then $e^T\tilde{Q}_1 e > 0$ and (B.7) gives $1 - \lambda > 0$ or $\lambda < 1$.

If on the other hand $\lambda < 1$ all λ then $e^T\tilde{Q}_1 e > 0$ for a basis of vectors and $\tilde{Q}_1 > 0$ gives $Q > 0$.

ii) By repeating the argument above and using $\geqslant$ instead of this part follows in the same way. The equivalence in part i) is to be used too.

iii) This part follows from i) and ii) by a simple negation.

Q.E.D.

Definition B.1:

$$\begin{bmatrix} x \\ y \end{bmatrix} = 0$$

is said to be a stable point if there is a $\delta > 0$ such that

$$\left|\left|\begin{bmatrix} x_o \\ y_o \end{bmatrix}\right|\right| < \delta$$

implies

$$\lim_{k\to\infty}\begin{bmatrix} x_k \\ y_k \end{bmatrix} = 0$$

With this use of stability it is easy to summarize the result.

Theorem B.1: Consider the function W given in (B.1).

i) If Q is positive definite, $x = 0$ is a stable point. The convergence is linear.

ii) If Q is indefinite, $x = 0$ is not a stable point.

Remark 1: Part ii) means that saddle points are of no interest when examining the possible limits of the minimization method.

Remark 2: Part i) means that the convergence is very slow. Close to a minimum point, the algorithm can be compared with the steepest descent method. The bound

$$\left[||x_{k+n}||^2 + ||y_{k+n}||^2\right]^{1/2} \leqslant$$

$$\leqslant [\max \lambda(P_1)]^n \left[||x_k||^2 + ||y_k||^2\right]^{1/2}$$

is easily obtained, where equality is possible.

Let for example $\max \lambda(P_1) = 0.95$, $||x_o|| = ||y_o|| = 1$, and $||x_n|| = ||y_n|| = 10^{-4}$. The bound above gives that 179 iterations are needed!

Since the eigenvalues of Q and P_1 vary continuously with the elements of Q it follows from Lemma B.5 that $\max \lambda(P_1) \rightarrow 1$ when the condition number of $Q \rightarrow \infty$.

APPENDIX C
ON CONDITIONS FOR LOCAL MINIMUM POINTS OF A SPECIAL FUNCTION.

In this appendix a special function is studied and its possible minimum points are examined. The reason for studying this function is that it can be interpreted as the loss function of the GLS method.

When the variance of the noise is small the equation $V' = 0$ will lead to equations of the type

$$f(x) + g(x) = 0, \quad g(x) = O(\varepsilon)$$

where ε is a small number. Some of the following lemmas deal with the properties of the solution of such equations.

The first lemma is the well-known principle of contraction mapping. It is stated here in order to later show how it can be used for the actual problems.

Lemma C.1: Let $B_\delta(x_o)$ denote the set $\{x;\ ||x-x_o|| \leqslant \delta\}$. Consider a map $S(x)$. If

i) $$||S(x_o) - x_o|| \leqslant (1-\alpha)\delta \qquad \alpha < 1 \tag{C.1}$$

ii) $$||S(x') - S(x")|| \leqslant \alpha||x' - x"||, \quad x',x" \in B_\delta(x_o) \tag{C.2}$$

then $S(x)$ has a unique fixpoint (a solution of $x = S(x)$) in $B_\delta(x_o)$.

Proof: See [17].

The next lemma deals with necessary properties of solutions. It does not guarantee existence or uniqueness of solutions.

Lemma C.2: Consider the equation

$$F(x,\varepsilon) = f(x) + g(x,\varepsilon) = 0 \tag{C.3}$$

where f and g are continuous functions.

Denote the null space of f by **Nf** $(Nf = \{x;\ f(x) = 0\})$
Let Ω be an arbitrary, compact set, which may depend on ε. Assume that

i) $\Omega - Nf$ is non empty

ii) there are constants $\varepsilon_1 > 0$ and $K < \infty$ such that $0 \leqslant \varepsilon \leqslant \varepsilon_1$ implies

$$\sup_{x \in \Omega} ||g(x,\varepsilon)|| \leqslant K\varepsilon$$

Then there is a number $\varepsilon_0 > 0$ such that if $0 \leqslant \varepsilon \leqslant \varepsilon_0$ and $\bar{x}$ is a solution of $F(x,\varepsilon) = 0$ then

$$\inf_{x_0 \in Nf} ||\bar{x} - x_0|| \to 0, \qquad \varepsilon \to 0 \tag{C.4}$$

Proof: Define a set $M(\varepsilon')$, a neighbourhood of Nf by

$$M(\varepsilon') = \{x;\ \inf_{x_0 \in Nf} ||x - x_0|| \leqslant \varepsilon'\}$$

By the construction and the continuity of f

$$\inf_{x \in \Omega - M(\varepsilon')} ||f(x)|| = \alpha(\varepsilon') > 0 \quad \text{if } \varepsilon' > 0$$

(where it is assumed that $\Omega - M(\varepsilon')$ is non empty).

Let $0 \leqslant \varepsilon \leqslant \varepsilon_1$. Then

$$\inf_{x \in \Omega - M(\varepsilon')} ||F(x,\varepsilon)|| \geq \inf_{x \in \Omega - M(\varepsilon')} ||f(x)|| -$$

$$- \sup_{x \in \Omega - M(\varepsilon')} ||g(x,\varepsilon)|| \geq \alpha(\varepsilon') - K\varepsilon$$

Define now $\varepsilon_o = \min \left(\varepsilon_1, \frac{1}{2}\frac{\alpha(\varepsilon')}{K}\right)$ which is strictly positive.

Let $0 \leq \varepsilon \leq \varepsilon_o$. Then

$$\inf_{x \in \Omega - M(\varepsilon')} ||F(x,\varepsilon)|| \geq \frac{1}{2}\alpha(\varepsilon') > 0$$

If $\bar{x}$ is a solution of (B.3) then $\bar{x} \in M(\varepsilon')$ and

$$\inf_{x_o \in Nf} ||\bar{x} - x_o|| \leq \varepsilon'$$

However, ε' can be chosen arbitrary small, so all solutions of (C.3) fulfil (C.4).

Q.E.D.

Corr: If $g(x,\varepsilon) = \varepsilon h(x,\varepsilon)$ where $h(x,\varepsilon)$ is a continuous function, the compact set Ω can be chosen arbitrarily.

The following lemma gives a sufficient condition for existence of a unique solution of the form (B.4).

Lemma C.3: Consider the equation

$$F(x,\varepsilon) = f(x) + g(x,\varepsilon) = 0 \tag{C.3}$$

where f and g are twice differentiable functions and $\dim f = \dim g = \dim x$.

Let x_o be a zero of $f(x)$ such that

i) $f'_x(x_o)$ is non singular

ii) there is a set $B_\delta(x_o) = \{x; ||x - x_o|| \leqslant \delta\}$ with δ (independent of ε) > 0, and constants ε_1, C_1 and C_2 such that

a) x_o is the only zero of $f(x)$ in $B_\delta(x_o)$,

b) $0 \leqslant \varepsilon \leqslant \varepsilon_1$ implies

$$\sup_{x \in B_\delta(x_o)} ||g(x,\varepsilon)|| \leqslant C_1\varepsilon$$

$$\sup_{x \in B_\delta(x_o)} ||g'_x(x,\varepsilon)|| \leqslant C_2\varepsilon$$

Then there is a number $\varepsilon_o > 0$ such that $0 \leqslant \varepsilon \leqslant \varepsilon_o$ implies

i) $F(x,\varepsilon) = 0$ has a unique solution $\bar{x}$ in $B_\delta(x_o)$

ii) $\bar{x}$ fulfils

$$\bar{x} - x_o = O(\varepsilon), \quad \varepsilon \to 0 \tag{C.5}$$

Proof: Study solutions of (C.3) in $B_{\delta_o}(x_o)$ where δ_o is an arbitrary constant satisfying $0 < \delta_o \leqslant \delta$.

Consider the function

$$S(x,\varepsilon) = x - f'_x(x_o)^{-1}F(x,\varepsilon)$$

If $S(x,\varepsilon)$ is a contraction mapping its fixpoint is the solution of $x - f'_x(x_o)^{-1}F(x,\varepsilon) = x$ of $F(x,\varepsilon) = 0$.
Put $C_o = ||f'_x(x_o)^{-1}||$.

Let $0 \leq \varepsilon \leq \varepsilon_1$. Then

$$||S(x_o,\varepsilon) - x_o|| \leq ||f'_x(x_o)^{-1}|| \cdot ||F(x_o,\varepsilon)|| \leq C_oC_1\varepsilon$$

Let x' and x" be two arbitrary, different points in $B_{\delta_o}(x_o)$. With use of the mean value theorem [17]

$$\frac{||S(x',\varepsilon) - S(x'',\varepsilon)||}{||x' - x''||} = \sup_{0\leq t\leq 1} ||S'_x(tx' + (1-t)x'',\varepsilon)||$$

Assume that the supremum is obtained at x = x'''.

$$\frac{||S(x',\varepsilon) - S(x'',\varepsilon)||}{||x' - x''||} \leq ||S'_x(x''',\varepsilon)|| =$$

$$= ||I - f'_x(x_o)^{-1}[f'_x(x''') + g'_x(x''',\varepsilon)]||$$

$$\leq C_o||f'_x(x''') - f'_x(x_o)|| + C_oC_1\varepsilon \leq C_oC_3\delta_o + C_oC_1\varepsilon$$

for some constants C_3 (depending on δ but not on δ_o).

Now (C.1) and (C.2) are fulfilled if

$$C_oC_1\varepsilon \leq (1-\alpha)\delta_o$$

$$C_oC_3\delta_o + C_oC_1\varepsilon \leq \alpha$$

Choose a value of α. Let δ_o satisfy

$$\delta_o = K\varepsilon$$

where

$$K \geq \frac{C_oC_1}{1-\alpha}$$

Define then

$$\varepsilon_o' = \min\left(\varepsilon_1, \frac{\delta}{K}, \frac{\alpha}{C_o(C_1 + C_3K)}\right) \tag{C.6}$$

Then (C.1), (C.2) and $\delta_o \leqslant \delta$ are fulfilled if $0 \leqslant \varepsilon \leqslant \varepsilon_o'$.

Now consider the set $\Omega = B_\delta(x_o) - B_{\delta_o}(x_o)$.

It has to be shown that $F(x,\varepsilon) = 0$ has no solutions in Ω if ε is small enough.

If δ_o is small enough

$$\inf_\Omega ||f(x)|| = \inf_{||x-x_o||=\delta_o} ||f(x)|| =$$

$$= \inf_{||x-x_o||=\delta_o} ||f(x_o) + f_x'(x_o)(x-x_o) +$$

$$+ O(||x-x_o||^2)|| = \alpha\delta_o + O(\delta_o^2)$$

α denotes the smallest singular value of $f_x'(x_o)$.

Thus there are constants ε_1' and C_4 such that $0 \leqslant \varepsilon \leqslant \varepsilon_1'$ implies

$$\inf_{x\in\Omega} ||F(x,\varepsilon)|| \geqslant \inf_{x\in\Omega} ||f(x)|| - \sup_{x\in\Omega} ||g(x,\varepsilon)|| \geqslant$$

$$\geqslant \alpha\delta_o - C_4\delta_o^2 - C_1\varepsilon$$

This expression should be positive. Insert $\delta_o = K\varepsilon$.

$$\varepsilon[(\alpha K - C_1) - C_4K^2\varepsilon] > 0$$

Now choose finally

$$K = \max\left(\frac{C_o C_1}{1-\alpha}, \frac{3C_1}{\alpha}\right)$$

and

$$\varepsilon_o = \min\left(\varepsilon_o', \varepsilon_1', \frac{C_1}{C_4 K^2}\right)$$

With these values of K and ε_o and with $\delta_o = K\varepsilon$ it can be seen by going through the proof once more that $F(x,\varepsilon) = 0$ has a unique solution in $B_{\delta_o}(x_o)$ and no solution in $B_\delta(x_o) - B_{\delta_o}(x_o)$.

Q.E.D.

Remark: If $f_x'(x_o)$ is singular, nothing general can be stated. Consider the scalar examples $F_1(x) = x^2 - \varepsilon$ and $F_2(x) = x^2 + \varepsilon$. $F_1(x)$ has zeros close to $x_o = 0$, but these do not satisfy (C.5). $F_2(x)$ has no real zeros at all.

Near a local extremum the matrix of second order derivatives plays a fundamental role for determining the character of the extremum. The following lemmas which deal with quadratic forms will be useful in the analysis of this matrix.

Lemma C.4: Consider the symmetric matrix

$$Q = \begin{bmatrix} A + \varepsilon A_1 & \varepsilon B \\ \varepsilon B^T & \varepsilon C \end{bmatrix} \tag{C.7}$$

and the vector

$$r = \begin{bmatrix} \varepsilon b \\ 0 \end{bmatrix} \tag{C.8}$$

Assume that A and C are positive definite. Then if $0 < \varepsilon \leqslant \varepsilon_o$ where $1/\varepsilon_o >$ the largest eigenvalue of $A^{-1}[A_1 - BC^{-1}B^T]$

i) Q is positive definite

ii) $Q^{-1}r = O(\varepsilon)$, $\varepsilon \rightarrow 0$

<u>Proof</u>:

i) By Lemma C.2 $Q > 0$ is equivalent to

$A + \varepsilon A_1 - \varepsilon B(\varepsilon C)^{-1}\varepsilon B^T > 0$ or

$$A + \varepsilon D > 0 \tag{C.9}$$

where $D = A_1 - BC^{-1}B^T$.

(C.9) is apparently true for small values of ε (since the eigenvalues of $A + \varepsilon D$ are continuous functions of ε). ε must only be smaller than the smallest number δ such that

$$\det[A + \delta D] = 0 \tag{C.10}$$

(C.10) is rewritten as

$$\det[A\delta(\frac{1}{\delta} I + A^{-1}D)] = \det(A\delta)\det(\frac{1}{\delta} I + A^{-1}D) = 0$$

From this equation it is seen that $1/\delta =$ the largest eigenvalue of $A^{-1}D$.

ii) Using formulas for the inverse of a partionated matrix [10]

$$Q^{-1}r = \begin{bmatrix} (A + \varepsilon D)^{-1}\varepsilon b \\ -C^{-1}B(A + \varepsilon D)^{-1}\varepsilon b \end{bmatrix}$$

If $\varepsilon < \delta$ then $[A + \varepsilon D]^{-1} = A^{-1} + O(\varepsilon)$ and $Q^{-1}r = O(\varepsilon)$ follows easily.

Q.E.D.

Lemma C.5: Consider the function

$$V(x,\varepsilon) = \frac{1}{2}\, x^T Q(\varepsilon)x + x^T r(\varepsilon) \qquad (C.11)$$

with

$$Q(\varepsilon) = \begin{bmatrix} A + \varepsilon A_1 & \varepsilon B \\ \varepsilon B^T & \varepsilon C \end{bmatrix} \qquad r(\varepsilon) = \begin{bmatrix} \varepsilon b \\ 0 \end{bmatrix}$$

with A_1 in a symmetric matrix, A and C are symmetric and positive definite matrices. There is a constant $\varepsilon_o >$ such that if $0 < \varepsilon \leqslant \varepsilon_o$ then:

To every $K_2 > 0$ there is a constant K_1 (depending on K_2 and ε_o but not on ε) such that

$$\inf_{||x||=K_1\varepsilon} V(x,\varepsilon) \geqslant K_2\varepsilon^2 \qquad (C.12)$$

Proof: Consider the set

$$\Omega(V_o,\varepsilon) = \{x;\ V(x,\varepsilon) \leqslant V_o\}$$

Define

$$x_o(\varepsilon) = -Q(\varepsilon)^{-1}r(\varepsilon)$$

Then $\Omega(V_o,\varepsilon)$ is given by

$$\frac{1}{2}\,(x - x_o(\varepsilon))^T Q(\varepsilon)(x - x_o(\varepsilon)) \leqslant V_o + \frac{1}{2}\,x_o(\varepsilon)^T Q(\varepsilon)x_o(\varepsilon) \quad (C.13)$$

$\Omega(V_o,\varepsilon)$ is non empty if

$$0 < \varepsilon < \delta$$

$$V_o \geqslant -\frac{1}{2}\,x_o(\varepsilon)^T Q(\varepsilon)x_o(\varepsilon)$$

where δ is the largest eigenvalue of $A^{-1}[A_1 - BC^{-1}B^T]$.

Let x_i denote the i:th component of x.
Define a new set

$$\Omega_1(V_o,\varepsilon) = \{x;\ |x_i - x_o(\varepsilon)_i| \leqslant \sup_{x \in \Omega(V_o,\varepsilon)} |x_i - x_o(\varepsilon)_i| \text{all } i\}$$

Clearly $\Omega(V_o,\varepsilon) \subseteq \Omega_1(V_o,\varepsilon)$.

What is $\sup\limits_{x \in \Omega(V_o,\varepsilon)} |x_i - x_o(\varepsilon)_i|$?

Let e_i denote a unit vector, which i:th component is 1.

Then the maximum of $e_i^T(x - x_o(\varepsilon))$ under the constraint

$$(x - x_o(\varepsilon))^T Q(\varepsilon)(x - x_o(\varepsilon)) = 2V_o + x_o(\varepsilon)^T Q(\varepsilon)x_o(\varepsilon)$$

is sought.

Using a Lagrange multiplier

$$e_i + \lambda 2Q(\varepsilon)\left(x - x_o(\varepsilon)\right) = 0$$

$$\left(x - x_o(\varepsilon)\right)^T Q(\varepsilon)\left(x - x_o(\varepsilon)\right)^T = 2V_o + x_o(\varepsilon)^T Q(\varepsilon) x_o(\varepsilon)$$

from which

$$\sup_{x \in \Omega(V_o,\varepsilon)} |x_i - x_o(\varepsilon)_i| = \sqrt{\frac{2V_o + x_o^T(\varepsilon)Q(\varepsilon)x_o(\varepsilon)}{[Q(\varepsilon)^{-1}]_{ii}}} \qquad (C.14)$$

is obtained by straight forward calculations.

The sphere

$$S_1(V_o,\varepsilon) = \left\{x;\ ||x - x_o(\varepsilon)|| \leq \right.$$

$$\left. \leq \sqrt{\sum_i \left[\frac{2V_o + x_o^T(\varepsilon)Q(\varepsilon)x_o(\varepsilon)}{[Q(\varepsilon)^{-1}]_{ii}}\right]^2}\right\}$$

contains the set $\Omega(V_o,\varepsilon)$ and so does the sphere

$$S_2(V_o,\varepsilon) = \left\{x;\ ||x|| \leq ||x_o(\varepsilon)|| + \right.$$

$$\left. + \sqrt{\sum_i \left[\frac{2V_o + x_o^T(\varepsilon)Q(\varepsilon)x_o(\varepsilon)}{[Q(\varepsilon)^{-1}]_{ii}}\right]^2}\right\}$$

A graphical illustration of the sets $\Omega(V_o,\varepsilon)$, $\Omega_1(V_o,\varepsilon)$, $S_1(V_o,\varepsilon)$ and $S_2(V_o,\varepsilon)$ for a two dimensional example is given in Fig. C.1.

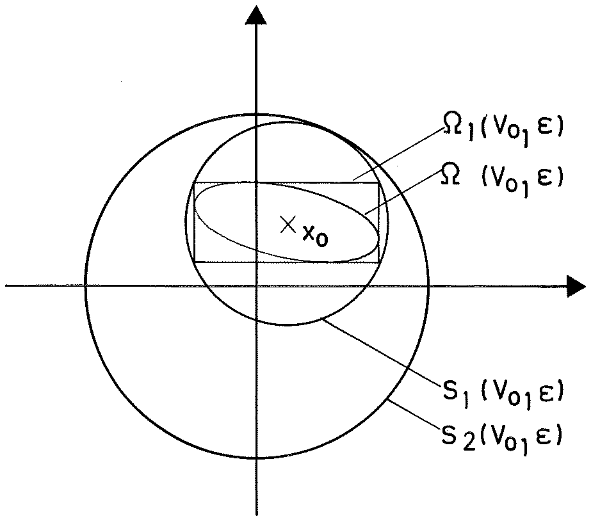


Fig. C.1.

The function $V(x,\varepsilon)$ has the following property. Let M_1 and M_2 be two convex and compact sets, containing $x_o(\varepsilon)$ and with boundaries ∂M_1 and ∂M_2. If $M_1 \subset M_2$ then

$$\inf_{x \in \partial M_1} V(x,\varepsilon) \leq \inf_{x \in \partial M_2} V(x,\varepsilon).$$

This is true since $V(x,\varepsilon)$ is a convex function. Define $\bar{x}_2 \in \partial M_2$ by

$$V(\bar{x}_2,\varepsilon) = \inf_{x \in \partial M_2} V(x,\varepsilon)$$

There is at least one point $\bar{x}_1 \in \partial M_1$ such that

$$\bar{x}_1 = tx_o(\varepsilon) + (1-t)\bar{x}_2 \qquad 0 \leq t \leq 1$$

so

$$\inf_{x \in \partial M_1} V(x,\varepsilon) \leqslant V(\bar{x}_1,\varepsilon) \leqslant tV\big(x_0(\varepsilon),\varepsilon\big) + (1-t)V(\bar{x}_2,\varepsilon) \leqslant$$

$$\leqslant V(\bar{x}_2,\varepsilon) = \inf_{x \in \partial M_2} V(x_2,\varepsilon)$$

Put now $M_1 = \Omega(V_o,\varepsilon)$ and $M_2 = S_2(V_o,\varepsilon)$.

Applying this property

$$\inf_{||x||=R(V_o,\varepsilon)} V(x,\varepsilon) \geqslant V_o \tag{C.15}$$

where

$$R(V_o,\varepsilon) = ||x_o(\varepsilon)|| + \left[\big(2V_o + x_o^T(\varepsilon)Q(\varepsilon)x_o(\varepsilon)\big)q(\varepsilon)\right]^{1/2} \tag{C.16}$$

$$q(\varepsilon) = \sum_i \frac{1}{[Q(\varepsilon)^{-1}]_{ii}} \tag{C.17}$$

There are constants ε_o, C_1, C_2 and C_3 (with $\varepsilon_o < \delta$ and C_1, C_2, C_3 independent of ε) such that $0 < \varepsilon \leqslant \varepsilon_o$ implies

$$||x_o(\varepsilon)|| \leqslant C_1\varepsilon$$

$$||x_o^T(\varepsilon)Q(\varepsilon)x_o(\varepsilon)|| \leqslant C_2\varepsilon^2$$

$$q(\varepsilon) \leqslant C_3$$

The last inequality follows from the expression for the inverse of a partionated matrix [10].

Define

$$R_1(\varepsilon,V_o) = C_1\varepsilon + [C_3(2V_o + C_2\varepsilon^2)]^{1/2} \qquad (C.18)$$

Let now $0 < \varepsilon \leqslant \varepsilon_o$. Then $R(\varepsilon,V_o) \leqslant R_1(\varepsilon,V_o)$ and from the property of $V(x,\varepsilon)$ described above

$$\inf_{||x||=R_1(\varepsilon,V_o)} V(x) \geqslant V_o$$

Now take $K_2 > 0$ arbitrary and put $V_o = K_2\varepsilon^2$.
Then $R_1(\varepsilon,V_o) = K_1\varepsilon$ with

$$K_1 = C_1 + [C_3(2K_2 + C_2)]^{1/2}$$

and the lemma is proved.

Q.E.D.

In the following theorem the results of the foregoing lemmas are applied to a function of special structure. It will later turn out that the loss function of the GLS method has this structure.

Theorem C.1: Consider the function

$$V(x,y,\varepsilon) = \frac{1}{2}\, x^TP(y)x + \varepsilon h(x,y) \qquad (C.19)$$

where $P(y)$ is a positive definite matrix for all y, twice differentiable with respect to y and $h(x,y)$ a twice differentiable function. ε is considered as a fix parameter.

Then there are necessary and sufficient conditions for local minimum points in an arbitrary compact set Ω.

There is a constant $\varepsilon_o > 0$ such that if $0 < \varepsilon \leq \varepsilon_o$ the following is true.

i) Every stationary point of $V(x,y,\varepsilon)$ in Ω fulfils

$$(x,y) = (0,y_o) + \big(O(\varepsilon),o(1)\big), \quad \varepsilon \to 0 \qquad (C.20)$$

where y_o is a solution of

$$h_y'(0,y) = 0 \qquad (C.21)$$

If (x,y) is a local minimum point it is necessary that $h_{yy}''(0,y_o)$ is positive definite or positive semidefinite.

ii) If y_o is a solution of (C.21) and $h_{yy}''(0,y_o)$ is positive definite then there exists a unique local minimum of the form (C.20), and the point will in fact satisfy

$$(x,y) = (0,y_o) + \big(O(\varepsilon),O(\varepsilon)\big), \quad \varepsilon \to 0 \qquad (C.22)$$

The matrix of second order derivatives is positive definite in the minimum point.

Proof: The equation $V' = 0$ turns out to be

$$\begin{bmatrix} P(y)x \\ \\ \dfrac{\partial}{\partial y}\left(\dfrac{1}{2}\ x^T P(y)x\right) \end{bmatrix} + \varepsilon \begin{bmatrix} h_x'(x,y) \\ \\ h_y'(x,y) \end{bmatrix} = 0 \qquad (C.23)$$

and the matrix of second order derivatives

$$V'' = \begin{bmatrix} V''_{xx} & V''_{xy} \\ V''_{yx} & V''_{yy} \end{bmatrix} =$$

$$= \begin{bmatrix} P(y) & \frac{\partial}{\partial y}[P(x)y] \\ \frac{\partial}{\partial y}[P(x)y]^T & \frac{\partial^2}{\partial y^2}\left[\frac{1}{2} x^T P(y) x\right] \end{bmatrix} +$$

$$+ \ \varepsilon \begin{bmatrix} h''_{xx}(x,y) & h''_{xy}(x,y) \\ h''_{yx}(x,y) & h''_{yy}(y,y) \end{bmatrix} \qquad (C.24)$$

The first part of (C.23) yields the necessary condition

$$||x|| = \varepsilon||P(y)^{-1}h'_x(x,y)|| \leqslant K\varepsilon \qquad (C.25)$$

where

$$K = \sup_{(x,y) \in \Omega} ||P(y)^{-1}h'_x(x,y)||$$

Apply Lemma C.2 to the second part of (C.23) putting

$$f(y) = h'_y(0,y)$$

$$g(y,\varepsilon) = \frac{1}{\varepsilon}\frac{\partial}{\partial y}\left(\frac{1}{2} x^T P(y) x\right) + h'_y(x,y) - h'_y(0,y)$$

Assume that (C.25) holds. Then there is a number $\varepsilon'_o > 0$ such that if $0 < \varepsilon \leqslant \varepsilon'_o$ the following condition is necessary

$$y - y_o = o(1), \qquad \varepsilon \to 0 \qquad (C.26)$$

where y_o is some solution of

$$h_y'(0,y) = 0$$

If (x,y) is a minimum point, it is necessary that V'' is positive definite or positive semidefinite. From this it follows that the same must be true to V_{yy}'' and further that there is a number ε_o'' such that $0 < \varepsilon \leqslant \varepsilon_o''$ implies the same condition for $h_{yy}''(0,y_o)$.

The first part of the theorem is proved.

If $h_{yy}''(0,y_o)$ is positive definite, it follows from Lemma C.3 that there is a number $\varepsilon_o''' > 0$ such that $0 < \varepsilon \leqslant$ $\leqslant \varepsilon_o'''$ implies that (C.26) can be replaced by

$$y = y_o + O(\varepsilon) \qquad \text{(C.27)}$$

When ε is small

$$V(x,y,\varepsilon) - V(0,y_o,\varepsilon) = [x^T \quad (y-y_o)^T] \begin{bmatrix} \varepsilon h_x'(0,y_o) \\ 0 \end{bmatrix} +$$

$$+ \frac{1}{2}[x^T \quad (y-y_o)^T] \begin{bmatrix} P(y_o) + \varepsilon h_{xx}(0,y_o) & \varepsilon h_{xy}(0,y_o) \\ \varepsilon h_{yx}(0,y_o) & \varepsilon h_{yy}(0,y_o) \end{bmatrix} \cdot$$

$$\cdot \begin{bmatrix} x \\ y-y_o \end{bmatrix} + r(x,y,\varepsilon)$$

where $r(x,y,\varepsilon) = O\left(||(x,y) - (0,y_o)||^3\right)$.

A straight forward application of Lemma C.5 gives: there are constants ε_o^{1V}, K_1 and K_2 such that $0 < \varepsilon \leqslant \varepsilon_o^{1V}$ implies

$$\inf_{||(x,y)-(0,y_o)||=K_1\varepsilon} V(x,y,\varepsilon) - V(0,y_o,\varepsilon) - r(x,y,\varepsilon) \geqslant K_2\varepsilon^2$$

But there are constants ε_o^{v} and K_3 such that $0 < \varepsilon \leqslant \varepsilon_o^{v}$ implies

$$\sup_{||(x,y)-(0,y_o)||=K_1\varepsilon} r(x,y,\varepsilon) \leqslant K_3\varepsilon^3$$

Thus

$$\inf_{||(x,y)-(0,y_o)||=K_1\varepsilon} V(x,y,\varepsilon) \geqslant V(0,y_o,\varepsilon) + K_2\varepsilon^2 - K_3\varepsilon^3$$

is greater than $V(0,y_o,\varepsilon)$ if $K_2 - K_3\varepsilon > 0$.

Put

$$\varepsilon_o^{vi} = \frac{K_2}{2K_3}$$

Then $0 < \varepsilon \leqslant \min(\varepsilon_o^{iv},\varepsilon_o^{v},\varepsilon_o^{vi})$ implies the existence of a local minimum point in the set

$$S(\varepsilon) = \{(x,y);\ ||(x,y) - (0,y_o)|| \leqslant K_1\varepsilon\}$$

When $(x,y) \in S(\varepsilon)$

$$V'' = \begin{bmatrix} P(y_o) + O(\varepsilon) & O(\varepsilon) \\ O(\varepsilon) & \varepsilon h''_{yy}(0,y_o) + O(\varepsilon^2) \end{bmatrix}$$

By Lemma C.4 it follows that there is a constant ε_o^{vii} such that $0 < \varepsilon \leqslant \varepsilon_o^{vii}$ and $(x,y) \in S(\varepsilon)$ imply that V'' is positive definite. From this it follows that $V(x,y,\varepsilon)$

has a unique minimum point in $S(\varepsilon)$.

Finally, choose $\varepsilon_o = \min(\varepsilon_o', \varepsilon_o'', \varepsilon_o''', \varepsilon_o^{iv}, \varepsilon_o^{v}, \varepsilon_o^{vi}, \varepsilon_o^{vii})$. Going through the proof once more, it is seen that all parts hold.

Q.E.D.

Remark 1: The greatest possible value of ε_o may depend on Ω. It is in general not possible to take Ω as the whole space. A simplified example: $V_\varepsilon(x) = x^2 + \varepsilon(x^3-x)$ has two stationary points: $x_1(\varepsilon) = -\frac{2}{\varepsilon} + O(\varepsilon)$ and $x_2(\varepsilon) = \frac{\varepsilon}{2} + O(\varepsilon^3)$ while $V_o(x)$ has one stationary point, $x = 0$.

Remark 2: If $h''_{yy}(0,y_o)$ is positive semidefinite (singular) nothing general can be stated. An illustrative example is

$$V(x,y) = \frac{1}{2}x^2 + \varepsilon\left[\frac{1}{2}x^2 + xy + Ky^n\right]$$

where the integer $n \geqslant 3$.

The equation (C.21) has the only solution $y = 0$ and $h''_{yy}(0,0) = 0$. For this function

$$V' = \begin{bmatrix} x + \varepsilon x + \varepsilon y \\ \varepsilon x + \varepsilon K n y^{n-1} \end{bmatrix}$$

$$V'' = \begin{bmatrix} 1+\varepsilon & \varepsilon \\ \varepsilon & \varepsilon K n(n-1) y^{n-2} \end{bmatrix}$$

The stationary points are the solutions of

$$x = -\frac{\varepsilon}{1+\varepsilon}\,y$$

$$y\left(y^{n-2} - \frac{\varepsilon}{(1+\varepsilon)Kn}\right) = 0$$

$(x,y) = (0,0)$ is always a stationary point and a saddle point. If

$$y^{n-2} = \frac{\varepsilon}{(1+\varepsilon)Kn}$$

has a solution then V" is positive definite in that point. This implies

i) If n is odd, there is one minimum point and $x = O(\varepsilon)$, $y = O\left(\varepsilon^{1/(n-2)}\right)$.

ii) If n is even and $K > 0$, there are two minimum points and $x = O(\varepsilon)$, $y = O\left(\varepsilon^{1/(n-2)}\right)$.

iii) If n is even and $K < 0$, there are no minimum points.

APPENDIX D.

ANALYSIS OF THE NOISE CONDITION (NC) FOR FIRST ORDER MODELS.

In order to prove Lemma 3.2 it is necessary to study the derivatives of V_2 and the solutions of $V_2' = 0$.

First order derivatives.

Direct computations give

$$\begin{cases} \frac{1}{2} V'_{\hat{a}} = (\hat{a} + \hat{c} + \hat{a}\hat{c}^2)r_o + (1 + 2\hat{a}\hat{c} + \hat{c}^2)r_1 + \hat{c}r_2 \\ \frac{1}{2} V'_{\hat{c}} = (\hat{a} + \hat{c} + \hat{a}^2\hat{c})r_o + (1 + 2\hat{a}\hat{c} + \hat{a}^2)r_1 + \hat{a}r_2 \end{cases} \tag{D.1}$$

The equations $V_2' = 0$ are rewritten

$$\begin{cases} (\hat{a} + \hat{c} + \hat{a}\hat{c}^2)r_o + (1 + 2\hat{a}\hat{c} + \hat{c}^2)r_1 + \hat{c}r_2 = 0 \\ [(\hat{a} - \hat{c})][\hat{a}\hat{c}r_o + (\hat{a} + \hat{c})r_1 + r_2] = 0 \end{cases} \tag{D.2}$$

Case i): A possible solution fulfils

$$\begin{cases} \hat{a} = \hat{c} \\ (2\hat{a} + \hat{a}^3)r_o + (1 + 3\hat{a}^2)r_1 + \hat{a}r_2 = 0 \end{cases} \tag{D.3}$$

Let $f(x) = (2x + x^3)r_o + (1 + 3x^2)r_1 + xr_2$.
With use of the relation

$$r_2 > - r_o + 2 \frac{r_1^2}{r_o}$$

which holds since w(t) is persistently exciting of order 3,

$$f(1) = 3r_o + 4r_1 + r_2 > \frac{2}{r_o}(r_o + r_1)^2 > 0$$

$$f(-1) = -3r_o + 4r_1 - r_2 < -\frac{2}{r_o}(r_o - r_1)^2 < 0$$

$$f'(x) = (2 + 3x^2)r_o + 6xr_1 + r_2 >$$

$$> \frac{1}{r_o}[(r_o^2 - r_1^2) + 3(xr_o + r_1)^2] > 0$$

From these inequalities it is concluded that (D.3) has a unique solution, which satisfies $|\hat{a}| < 1$.

Case ii): The other possibility can be written

$$\begin{cases} (\hat{a} + \hat{c})r_o + (1 + \hat{a}\hat{c})r_1 = 0 \\ \hat{a}\hat{c}r_o + (\hat{a} + \hat{c})r_1 + r_2 = 0 \end{cases} \qquad (D.4)$$

Introducing the new variables $\hat{d}_1 = \hat{a} + \hat{c}$, $\hat{d}_2 = \hat{a}\hat{c}$ it is found that $\hat{a}$ and $\hat{c}$ are the roots of

$$z^2 - \hat{d}_1 z + \hat{d}_2 = 0 \qquad (D.5)$$

$$\begin{bmatrix} r_o & r_1 \\ r_1 & r_o \end{bmatrix} \begin{bmatrix} \hat{d}_1 \\ \hat{d}_2 \end{bmatrix} + \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = 0 \qquad (D.6)$$

Real valued solutions of (D.5) exist when the discriminant $\hat{d}_1^2 - 4\hat{d}_2 \geq 0$ or invoking (D.6)

$$D^* = r_1^2(r_2 - r_o)^2 - 4(r_o^2 - r_1^2)(r_1^2 - r_o r_2) \geq 0 \qquad (D.7)$$

Proof of Lemma 3.2: From the analysis above it is clear that

i) if $D^* < 0$ then $V_2' = 0$ has one solution

ii) if $D^* = 0$ then $V_2' = 0$ has three coincident solutions

iii) if $D^* > 0$ then $V_2' = 0$ has three different solutions.

Only the case $D^* > 0$ has to be considered closer.

The change of variables means that the function

$$E[\hat{D}(q^{-1})w(t)]^2, \qquad \hat{D}(q^{-1}) = 1 + \hat{d}_1 q^{-1} + \hat{d}_2 q^{-2}$$

is minimized. This function has a unique minimum with a positive definite matrix of second order derivatives.

When $D^* > 0$ the solutions of (D.5) satisfy $\hat{a} \neq \hat{c}$ and the Jacobian of the transformations of variables is non singular. This fact implies that V_2'' is positive definite for solutions of (D.5) if $D^* > 0$.

Q.E.D.

APPENDIX E.
PROOF OF THEOREM 3.4.

In this appendix it will be shown that by changing variables, Theorem 3.4 follows from Theorem C.1.

Proof of Theorem 3.4: Introduce the vectors (as in the proof of Theorem 3.2)

$$x = \begin{bmatrix} \hat{a}_1 - a_1 \\ \vdots \\ \hat{a}_n - a_n \\ \vdots \\ \hat{a}_{n+k} \\ \hat{b}_1 - b_1 \\ \vdots \\ \hat{b}_n - b_n \\ \vdots \\ \hat{b}_{n+k} \end{bmatrix} \qquad y = \begin{bmatrix} \hat{c}_1 \\ \vdots \\ \hat{c}_{n+k} \end{bmatrix} \tag{E.1}$$

The loss function can be written

$$V(x,y) = \frac{1}{2} x^T P(y) x + \varepsilon h(x,y) \tag{E.2}$$

with P(y) as the covariance matrix of the system

$$A(q^{-1})y^F(t) = - B(q^{-1})u^F(t), \qquad u^F(t) = \hat{C}(q^{-1})u(t)$$

P(y) is, however, always singular, but the null space of P(y) is independent of y. This is obvious, since from Theorem 2.2 the null space is spanned by vectors of the form

$$\begin{bmatrix} f_1 \\ \vdots \\ f_{n+k} \\ g_1 \\ \vdots \\ g_{n+k} \end{bmatrix} \qquad (E.3)$$

with

$$F(q^{-1}) = \sum_{i=1}^{n+k} f_i q^{-i} = A(q^{-1})L'(q^{-1}) \qquad (E.4)$$

$$G(q^{-1}) = \sum_{i=1}^{n+k} g_i g^{-i} = B(q^{-1})L'(q^{-1}) \qquad (E.5)$$

$$L'(q^{-1}) = \sum_{i=1}^{k} \ell_i' q^{-i} \text{ arbitrary} \qquad (E.6)$$

Introduce now the new variables

$$x' = \begin{bmatrix} x_1' \\ x_2' \end{bmatrix}$$

where x_1' is of dimension k and x_2' of dimension 2n+k. The vector x' is defined by

$$x = Qx' = [Q_1 \mid Q_2] \begin{bmatrix} x_1' \\ x_2' \end{bmatrix} \qquad (E.7)$$

where

$$Q_1 = \left[\begin{array}{ccc}
1 & & 0 \\
a_1 & \ddots & \\
\vdots & & \\
a_n & & 1 \\
 & \ddots & \vdots \\
0 & & a_n \\
\hline
0 & & 0 \\
b_1 & & \\
\vdots & \ddots & \\
b_n & & b_1 \\
 & \ddots & \vdots \\
0 & & b_n
\end{array}\right] \qquad \text{(E.8)}$$

Q_1 is a (2n+2k) × k matrix and Q an arbitrary (2n+2k) × × (2n+k) matrix with the properties $Q_1^T Q_2 = 0$ and Q non singular. Q_2 can for instance be constructed by Gram Schmidt orthogonalization.

From the discussion it follows that

$Q_1 x_1'$ is a typical element in the null space $N\big(P(y)\big)$
$Q_2 x_2'$ is a typical element in the space $N\big(P(y)\big)^{\perp}$

From these facts it is concluded that

$$P(y)Q_1 = 0$$

and that the matrix

$$R(y) = Q_2^T P(y) Q_2 \qquad \text{(E.9)}$$

of order (2n+k) × (2n+k) is non singular for all y.

The loss function is now written as

$$V(x_2',z) = \frac{1}{2}\, x_2'^T R(z) x_2' + \varepsilon k(x_2',z) \tag{E.10}$$

where z denotes the vector

$$\begin{bmatrix} x_1' \\ y \end{bmatrix}$$

Write the vector x_1' as

$$x_1' = \begin{bmatrix} \ell_1 \\ \vdots \\ \ell_k \end{bmatrix} \tag{E.11}$$

Then $x = Q_1 x_1'$ is equivalently expressed as

$$\hat{A}(q^{-1}) = A(q^{-1})\hat{L}(q^{-1}), \quad \hat{B}(q^{-1}) = B(q^{-1})\hat{L}(q^{-1}) \tag{E.12}$$

with

$$L(q^{-1}) = 1 + \hat{\ell}_1 q^{-1} + \dots + \hat{\ell}_k q^{-k} \tag{E.13}$$

The function k(0,z) is written by operators as

$$k(0,z) = E[\hat{L}(q^{-1})\hat{C}(q^{-1})v(t)]^2$$

Invoking Theorem C.1 the proof is finished.

Q.E.D.

APPENDIX F.

CONSTRUCTION OF COUNTER EXAMPLES TO THE SECOND VERSION OF GLS.

The equations (3.34) - (3.64) for the example of Section 3.8 are examined in this appendix.

(3.36) has the solution

$$\hat{a} = - \frac{r_y(1)}{r_y(0)}$$

$$\hat{b} = 1$$

from which

$$\varepsilon(t) = \frac{1 + \hat{a}q^{-1}}{1 + aq^{-1}} \cdot q^{-1}u(t) + \frac{1 + \hat{a}q^{-1}}{1 + aq^{-1}} v(t)$$

Define the functions F, f and g by

$$F(a,c) = r_\varepsilon(1) = f(a,c) + Sg(a,c)$$

$$f(a,c) = E\left[\frac{1 + \hat{a}'q^{-1}}{1 + aq^{-1}} v(t) \cdot \frac{1 + \hat{a}'q^{-1}}{1 + aq^{-1}} v(t+1)\right]$$

with

$$\hat{a}' = - \frac{r_y'(1)}{r_y'(0)}, \qquad y'(t) = \frac{1}{1 + aq^{-1}} v(t)$$

g(a,c) is a differentiable function.

Consider now especially

$$v(t) = \frac{1}{1 + cq^{-1}} e(t)$$

Then

$$\hat{a}' = \frac{a + c}{1 + ac} \quad \text{and} \quad f(a,c) = \frac{-ac(a+c)}{(1-ac)(1+ac)^2}$$

Further

$$f(0,c) = 0 \qquad f_a'(0,c) = \frac{-c^2}{(1-c)(1+c)^2}$$

$$f(-c,c) = 0 \qquad f_a'(-c,c) = \frac{c^2}{(1-c^2)^2(1+c^2)}$$

if $c \neq 0$ the existence of solutions of the forms

$a = O(S)$
$a = - c + O(S)$

now follow from Lemma C.3.

APPENDIX G.
DESCRIPTION OF PROGRAMS.

The main structure of the program package for the GLS identification is given in the table below. In the following pages a more detailed description of every subroutine is given.

Program or subroutine	Purpose	Called subroutines
TGLS	Main program	SIMUL GLS
SIMUL	Simulates the system	PRBSTA PRB NODI
GLS	Performs the GLS identification	LS FILT RESID VGLS
PRBSTA, PRB	Generates a PRBS	-
NODI	Generates white noise	-
LS	Performs a LS identification	LSQ
LSQ	Computes a least squares solution	-
FILT	Filters data	-
RESID	Computes the residuals	-
VGLS	Computes the loss function and related variables	FILT DSYMIN EIGS
DSYMIN	Invertes a symmetric matrix	-
EIGS	Computes eigenvalues and eigenvectors of a symmetric matrix	-

In subroutine VGLS there is a possibility to improve the solution by making some (approximative) Newton Raphson iterations.

```
C     PROGRAM TGLS
C
C     MAIN PROGRAM FOR GENERALIZED LEAST SQUARES IDENTIFICATION
C     OF SIMULATED DATA
C     AUTHOR TORSTEN SODERSTROM 1971-10-01
C
C     THE FOLLOWING DATA ARE READ FROM CARDS
C  1- M,ISYST,IMOD,INF - 4I10
C     M - NUMBER OF SAMPLES (MAX 1000)
C     IF M=0 THE PROGRAM STOPS
C     ISYST=10000*NA+100*NB+NC  ORDER OF TRUE OPERATORS
C     IMOD=10000*MNA+100*MNB+MNC  ORDER OF ESTIMATED OPERATORS
C     INF=10000*ITER+1000*IFILT+100*INIT+10*IPRINT+ISIM
C     ITER - MAX NUMBER OF ITERATIONS
C     IFILT  =0-FILTER ORIGINAL DATA   =1-FILTER FILTERED DATA
C     INIT  =0 START WITH LS ESTIMATE OF A AND B
C           =1 START WITH VALUES OF A AND B FROM CARD
C           =2 START WITH VALUES OF C       FROM CARD
C     IPRINT  =0-LITTLE OUTPRINT  =1 GREAT OUTPRINT
C     ISIM =0 U(T) IS A PRBS
C          =1 U(T) IS A WHITE NOISE INDEPENDENT OF E(T)
C  2- (T(I),I=1,(NA+NB+NC)),AL -  8F10.5
C     T - PARAMETER VECTOR  (TRUE VALUES)
C     AL - STANDARD DEVIATION OF THE NOISE
C  3- /IF INIT=1/ (T(I),I=1,(MNA+MNB)) - 8F10.5 START VALUES OF A AND B
C  3- /IF INIT=2/ (T(I)+MNA+MNB),I=1,MNC) -8F10.5  START VALUES OF C
C     SUBROUTINE REQUIRED
C            SIMUL
C            PRBSTA
C            PRB
C            NODI
C            GLS
C            LS
C            LSQ
C            VGLS
C            DSYMIN
C            FILT
C            RESID
C            EIGS
C
      DIMENSION U(1000),Y(1000),DAT(3000),AB(1000,11)
      DIMENSION TSYST(30),TMOD(30)
C
```

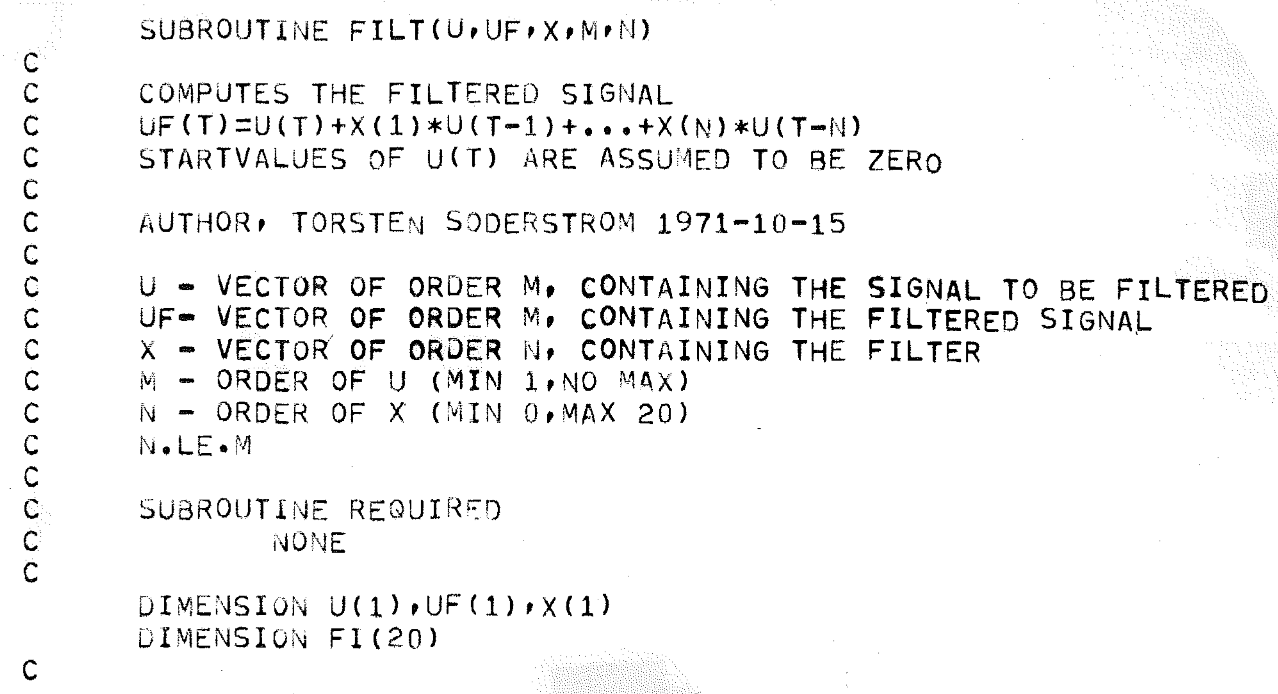
```
      SUBROUTINE SIMUL(U,Y,T,AMPL,AL,M,NA,NB,NC,IU1,IU2,IE)
C
C     COMPUTES A SIMULATION OF THE SYSTEM
C     A(Q)*Y(T)=B(Q)*U(T)+C(Q)*E(T)     E(T) GAUSSIAN WHITE NOISE
C     A(Q)=1 + A(1)*Q**(-1) +...+ A(NA)*Q**(-NA)
C     B(Q)=    B(1)*Q**(-1) +...+ B(NB)*Q**(-NB)
C     C(Q)=1 + C(1)*Q**(-1) +...+ C(NC)*Q**(-NC)
C     STARTVALUES OF U(T) AND E(T) ARE ZERO.
C     AUTHOR TORSTEN SODERSTROM, 1971-10-01
C
C     U - VECTOR OF ORDER M CONTAINING THE INPUT
C     Y - VECTOR OF ORDER M CONTAINING THE OUTPUT
C     T - VECTOR OF ORDER (NA+NB+NC) CONTAINING THE PARAMETERS
C     T=(A(1)...A(NA),B(1)...B(NB),C(1)...C(NC))
C     AMPL - AMPLITUDE OR STANDARD DEVIATION OF THE INPUT SIGNAL
C     AL - STANDARD DEVIATION OF THE NOISE
C     M - ORDER OF U,Y  (MIN 1,NO MAX)
C     NA - ORDER OF A
C     NB - ORDER OF B
C     NC - ORDER OF C
C     (NA+NB+NC) (MIN 0,MAX 30)
C     IU1 - =1 THE INPUT SIGNAL IS A PRBS.
C            2 THE INPUT SIGNAL IS A STEP AT TIME T=1
C            3 THE INPUT SIGNAL IS AN IMPULSE AT TIME T=1
C            4 THE INPUT SIGNAL IS WHITE NOISE INDEPENDENT OF E(T)
C            5 THE INPUT SIGNAL IS CONTAINED IN U
C     IU2 - NUMBER OF BITS IN THE SHIFTREGISTER FOR THE PRBSGENERATOR
C     (MIN3,MAX17)
C     IE - STARTVALUES TO NODI  IE MUST BE AN ODD INTEGER
C
C     ATTENTION. FOR BEST RESULT THE VALUE OF IE MUST BE CHOSEN WITH CARE
C
C     SUBROUTINE REQUIRED
C             NODI
C             PRBSTA
C             PRB
C
      DIMENSION U(1),Y(1),T(1)
      DIMENSION FI(30),LA(17),LX(17)
C
```

```
      SUBROUTINE GLS(DAT,T,AB,M,NA,NB,NC,ITER,ITER1,IFILT,INIT,IPRINT,
     FEPST,IA,IB)
C
C     COMPUTES THE GENERALIZED LEAST SQUARES ESTIMATE
C     A(Q)*C(Q) Y(T) = B(Q)*C(Q) U(T) + E(T)
C     A(Q)=1 + A(1)*Q**(-1) +...+ A(NA)*Q**(-NA)
C     B(Q)=    B(1)*Q**(-1) +...+ B(NB)*Q**(-NB)
C     C(Q)=1 + C(1)*Q**(-1) +...+ C(NC)*Q**(-NC)
C
C     AUTHOR  TORSTEN SODERSTROM  1971-10-01
C     DAT - VECTOR OF ORDER 3*M, CONTAINING THE DATA IN THE FOLLOWING FORM
C     TIME(1),U(1),Y(1),TIME(2),... Y(M)
C     T - VECTOR OF ORDER (NA+NB+NC) AT RETURN CONTAINING THE PARAMETER
C     ESTIMATES
C     T = (A(1),...A(NA),B(1),...B(NB.,C(1),...C(NC))
C     AB - MATRIX OF ORDER M*(NA+NB+NC) USED INTERNLY
C     M - ORDER OF U AND Y (NUMBER OF SAMPLES) (MIN 31,MAX 1000)
C     NA,NB,NC - ORDER OF A,B,C RESP.
C     (NA+NB+NC)  (MIN 0,MAX 30)
C     ITER - MAX NUMBER OF ITERATIONS (MIN 0,NO MAX)
C     ITER1 - MAX NUMBER OF VGLS-CALLS (MIN 1,NO MAX)
C     IFILT - IFILT=0 THE FILTER C(Q) IS APPLIED TO ORIGINAL DATA
C           - IFILT=1 THE FILTER C(Q) IS APPLIED TO FILTERED DATA
C     INIT - INIT=0 THE ITERATION IS STARTED WITH THE LS-ESTIMATES OF A AND B
C            INIT=1 THE ITERATION IS STARTED WITH GIVEN VALUES OF A AND B
C            INIT=2 THE ITERATION IS STARTED WITH GIVEN VALUES OF C
C     IPRINT - IPRINT =0 MINIMAL RESULTS ARE PRINTED
C              IPRINT =1 MEDIUM  RESULTS ARE PRINTED
C              IPRINT =2 MUCH    RESULTS ARE PRINTED
C     EPST - TEST QUANTITY FOR STOP OF ITERATIONS
C     IA,IB DIMENSION PARAMETERS OF AB
C
C     THE VECTOR DAT IS NOT DESTROYED
C
C     SUBROUTINE REQUIRED
C            LS
C            LSQ
C            RESID
C            FILT
C            VGLS
C            DSYMIN
C            EIGS
C
      DIMENSION DAT(1),T(1),AB(IA,IB)
      DIMENSION U(1000),UF(1000),Y(1000),YF(1000),RES(1000),DATA(3000)
      DIMENSION T1(30),T2(30),TT(30),NNB(1)
      COMMON/LSCOM/ V,SS,P(50,50),C(50),Q(50)
C
```

```
      SUBROUTINE PRBSTA(LA,NA)
C
C     SUBROUTINE TO START UP THE PRB-SUBROUTINE
C
C     REFERENCES, W. W. PETERSON, ERROR-CORRECTING CODES
C     B. ROSENGREN AND I. NORDH, KONSTRUKTION AV PRBS-GENERATOR
C     M. RUDEMO, ON PSEUDO-RANDOM NOISE GENERATED BY SHIFT REGISTERS
C     AUTHOR, STURE LINDAHL 1970-02-10
C     REVISE, STURE LINDAHL 1970-11-24
C
C     LA VECTOR, CONTAINING THE FEEDBACK-POLYNOMIAL
C     NA NUMBER OF BITS IN THE SHIFTREGISTER
C
C     NA MUST BE IN THE RANGE 3.LE.NA.LE.17
C
C     SUBROUTINE REQUIRED
C            NONE
      DIMENSION LA(1)
      SUBROUTINE PRB(LA,LX,Y,NA,AMP)
C
C     SUBROUTINE TO GENERATE A NEW STATE IN A PRBS-GENERATOR
C     REFERENCES, W. W. PETERSON, ERROR CORRECTING CODES
C     B. ROSENGREN AND I. NORDH, KONSTRUKTION AV PRBS-GENERATOR
C     M. RUDEMO, ON PSEUDO-RANDOM NOISE GENERATED BY SHIFT REGISTERS
C     AUTHOR, STURE LINDAHL 1970-02-10
C     REVISED, STURE LINDAHL 1970-11-23
C
C     LA VECTOR, CONTAINING THE FEEDBACK-POLYNOMIAL
C     LX VECTOR, CONTAINING THE ACTUAL STATE
C     Y  OUTPUT FROM PRBS-GENERATOR
C     NA NUMBER OF BITS IN THE SHIFTREGISTER
C     AMP SPECIFIED AMPLITUDE OF OUTPUT-SIGNAL
C
C     LA CAN BE ASSIGNED VALUES IN A STARTROUTINE PRBSTA
C
C     SUBROUTINE REQUIRED
C            NONE
C
      DIMENSION LA(1),LX(1)
      SUBROUTINE NODI(NODD,GAUSS)
C
C      GENERATES RANDOM NUMBERS N(0,1).SUITED FOR REPEATED USE.
C     REFERENCE B JANSON, RANDOM NUMBER GENERATORS.
C     AUTHOR K EKLUND 9/9 1970
C
C
C     GAUSS-RETURNED CONTAINING A RANDOM NUMBER N(0,1)
C     NODD -BY FIRST CALL OF NODI, NODD MUST EQUAL AN ODD INTEGER
C           NODD IS RETURNED CONTAINING A NEW ODD INTEGER WHICH
C           IS USED BY REPEATED CALLS
C
C     SUBROUTINES REQUIRED
C         NONE
C
```

```
      SUBROUTINE LS(DAT,T,AB,M,NU,NA,NB,IA,IB,IPRINT)
C
C     COMPUTES LEAST SQUARES MODEL
C     Y(T)+A(1)*Y(T-1)+..+A(NA)*Y(T-NA)=
C     B1(1)*U1(T-1)+...B1(NB(1))*U1(T-NB(1))+...
C     BNU(1)*UNU(T-1)+...BNU(NB(NU))*UNU(T-NB(NU)))+E(T)
C     AUTHOR, TORSTEN SODERSTROM, 1970-03-03
C     REVISED, TORSTEN SODERSTROM, 1971-10-01
C
C     DAT-VECTOR OF ORDER M*(NA+NB(1)+...+NB(NU)+1)
C     CONTAINING THE DATA IN THE FOLLOWING FORM
C     TIME(1),U1(1),U2(1),...UNU(1),Y(1)...
C     TIME(2),U1(2),U2(2),...UNU(2),Y(2),...
C     TIME(M),U1(M),U2(M),...UNU(M),Y(M)
C     T-VECTOR OF ORDER (NA+NB(1)+...NB(NU))
C     T=(A(1),..A(NA),B1(1),..B1(NB(1),B2(1),..BNU(NB(NU)))
C     AB-MATRIX OF ORDER M*(NA+NB(1)+...+NB(NU)+1) USED INTERNLY
C     M-NUMBER OF SAMPLES (NO MAX)
C     NA-NUMBER OF A-PARAMETERS.
C     NU-NUMBER OF INPUTS
C     NB-VECTOR OF ORDER NU
C     NB(I) IS THE NUMBER OF BI-PARAMETERS
C     THE FOLLOWING RESTRICTIONS ON M,NA,NU,NB MUST HOLD
C     (NA+NB(1)+...NB(NU)) (MIN 0,MAX 50)
C     NA+NB(1)+...NB(NU)+MAX(NA,NB(1),...NB(NU)) .LT. M
C     IA,IB - DIMENSION PARAMETERS OF AB
C     IPRINT-PRINT PARAMETER.
C     IPRINT=0-NOTHING IS PRINTED.
C     IPRINT=1 THE PARAMETERS ESTIMATES AND STANDARD DEVIATIONS
C              THE LOSS FUNCTION AND THE SINGULAR VALUES ARE PRINTED
C     IPRINT=2 AS IPRINT=1 + THE COVARIANCE MATRIX OF THE PARAMETER
C              ESTIMATES IS PRINTED
C
C     THE FOLLOWING VARIABLES LIE IN A COMMON BLOCK CALLED /LSCOM/
C     V-THE LOSS FUNCTION
C     S-ESTIMATED STANDARD DEVIATION OF THE NOISE
C     P-MATRIX OF DIMENSION 50*50 - THE COVARIANCE MATRIX OF
C     THE PARAMETER ESTIMATES
C     C-VECTOR OF DIMENSION 50 - THE STANDARD DEVIATION OF
C     THE PARAMETER ESTIMATES
C     Q-VECTOR OF DIMENSION 50 CONTAINING THE SINGULAR VALUES
C
C     THE VECTOR DAT IS NOT DESTROYED
C
C     SUBROUTINE REQUIRED
C             LSQ
C
      DIMENSION AB(IA,IB)
      DIMENSION DAT(1),T(1),NB(1)
      COMMON /LSCOM / V,S,P(50,50),C(50),Q(50)
      DIMENSION XX(50,1)
C
```

```
      SUBROUTINE LSQ(AB,XX,Q,EPS,MM,NN,JJP,IM,IN,IP,INP)
C
C     COMPUTES THE LEAST SQUARES SOLUTION  OF THE SYSTEM A*X=B USING
C     SINGULAR VALUE DECOMPOSITION.
C     REFERENCE, GOLUB-REINSCH,SINGULAR VALUE DECOMPOSITION AND
C     LEAST SQUARES SOLUTIONS.
C     AUTHOR,TORSTEN SODERSTROM,11/06-70.
C
C     AB-MATRIX OF ORDER MM*(NN+JJP). THE FIRST NN COLUMNS CONTAIN
C     THE MATRIX A.  THE LAST JJP COLUMMNS CONTAIN THE MATRIX B.
C     XX-MATRIX OF ORDER NN*JJP,RETURNED CONTAINING THE LEAST
C     SQUARES SOLUTION.
C     Q-VECTOR OF ORDER NN, RETURNED CONTAINING THE SINGULAR VALUES OF A.
C     EPS-IF ANY ELEMENT OF Q IS .LT. EPS*MAX Q(I), IT IS
C     CONSIDERED AS ZERO.
C     MM-NUMBER OF ROWS OF A (NO MAX).
C     NN-NUMBER OF COLUMNS OF A (MAX 50). NN .LE. MM.
C     JJP-NUMBER OF COLUMNS OF B (NO MAX).
C     IM,IN,IP,INP-DIMENSION PARAMETERS.
C
C     ATTENTION. THE MATRIX AB IS DESTROYED.
C
C     SUBROUTINE REQUIRED
C             NONE
C
      DIMENSION AB(IM,INP),XX(IN,IP),Q(IN)
      DIMENSION E(50)
C
```



```
      SUBROUTINE FILT(U,UF,X,M,N)
C
C     COMPUTES THE FILTERED SIGNAL
C     UF(T)=U(T)+X(1)*U(T-1)+...+X(N)*U(T-N)
C     STARTVALUES OF U(T) ARE ASSUMED TO BE ZERO
C
C     AUTHOR, TORSTEN SODERSTROM 1971-10-15
C
C     U - VECTOR OF ORDER M, CONTAINING THE SIGNAL TO BE FILTERED
C     UF- VECTOR OF ORDER M, CONTAINING THE FILTERED SIGNAL
C     X - VECTOR OF ORDER N, CONTAINING THE FILTER
C     M - ORDER OF U (MIN 1,NO MAX)
C     N - ORDER OF X (MIN 0,MAX 20)
C     N.LE.M
C
C     SUBROUTINE REQUIRED
C           NONE
C
      DIMENSION U(1),UF(1),X(1)
      DIMENSION FI(20)
C
```

```
      SUBROUTINE RESID(U,Y,RES,X,M,NA,NB)
C
C     COMPUTES THE RESIDUALS
C     RES(T)=Y(T)+A(1)*Y(T-1)+...A(NA)*Y(T-NA)-
C     -B(1)*U(T-1)-...-B(NB)*U(Y-NB)
C     RES(T)=0  T=1,... MAX(NA,NB)
C
C     AUTHOR  TORSTEN SODERSTROM 1971-10-15
C
C     U - VECTOR OF ORDER M, CONTAINING THE INPUT SIGNAL
C     Y - VECTOR OF ORDER M, CONTAINING THE OUTPUT SIGNAL
C     RES - VECTOR OF ORDER M ,CONTAINING THE RESIDUALS
C     X - VECTOR OF ORDER (NA+NB)
C     X=(A(1),...A(NA),B(1),...B(NB))
C     M- NUMBER OF SAMPLES  (MIN 1,NO MAX)
C     NA,NB - ORDER OF A RESP B
C     (NA+NB) (MIN 0,MAX 20)
C     MAX(NA,NB) .LT. M
C
C     SUBROUTINE REQUIRED
C            NONE
C
      DIMENSION U(1),Y(1),RES(1),X(1)
      DIMENSION FI(21)
C
```

```
      SUBROUTINE VGLS(U,UF,Y,YF,RES,T,M,NA,NB,NC,IFILT,IPRINT,ITMAX)
C
C     COMPUTES THE LOSS FUNCTION ETC FOR THE GLS PROBLEM
C
C     AUTHOR  TORSTEN SODERSTROM 1971-10-01
C
C     U - VECTOR OF ORDER M CONTAINING THE INPUT
C     UF- VECTOR OF ORDER M CONTAINING THE FILTERED INPUT
C     Y - VECTOR OF ORDER M CONTAINING THE OUTPUT
C     YF- VECTOR OF ORDER M CONTAINING THE FILTERED OUTPUT
C     RES-VECTOR OF ORDER M CONTAINING THE RESIDUALS  RES(T)=A(Q)*Y(T)-B(Q)*U(T)
C     T - VECTOR OF ORDER (NA+NB+NC) CONTAINING THE ACTUAL PARAMETER VALUES
C     M-ORDER OF U AND Y (NUMBER OF SAMPLES) (MIN 31,MAX 1000)
C     NA,NB,NC - NUMBER OF A,B,C PARAMETERS RESP
C     (NA+NB+NC) (MIN 0,MAX 30)
C     IFILT - IFILT=0 THE FILTER C(Q) IS APPLIED TO ORIGINAL DATA
C           - IFILT=1 THE FILTER C(Q) IS APPLIED TO FILTERED DATA
C     IPRINT -PRINT PARAMETER
C     THE FOLLOWING VARIABLES ARE PRINTED
C     IPRINT=0 THE LOSS FUNCTION AND THE GRADIENT
C              STANDARD DEVIATIONS OF THE PARAMETERS AND THE NOISE
C              EXTRAPOLATED PARAMETER ESTIMATES BASED ON NEWTON-RAPHSON
C     ELSE     AS IPRINT=0 +
C              THE MATRIX OF SECOND ORDER DERIVATIVES
C              ITS EIGENVALUES AND EIGENVECTORS
C              THE ESTIMATED COVARIANCE MATRIX OF THE PARAMETER ESTIMATES
C     ITMAX - MAX NUMBER OF NEWTON RAPHSON STEPS.
C
C     SUBROUTINE REQUIRED
C            FILT
C            RESID
C            DSYMIN
C            EIGS
C
      DIMENSION U(1),UF(1),Y(1),YF(1),RES(1),T(1)
      DIMENSION RESF(1000),VT(30),VTT(30,30),P(30,30),DT(30),
     FT2(30),R(30,30),EV(30),C(20)
      DOUBLE PRECISION P
C
```

```
      SUBROUTINE DSYMIN(N,IA,IFAIL,A)
C
C     DOUBLE PRECISION VERSION OF SUBROUTINE SYMIN.
C     SUBROUTINE FOR INVERSION OF SYMMETRIC MATRICES.
C     REFERENCE,RUTISHAUSER,CACM,ALG.NR.150.
C     AUTHOR,K.MORTENSSON 04/04-68.
C
C     A-MATRIX TO BE INVERTED.UPON RETURN A CONTAINS A-1 IF THE
C     INVERSION HAS SUCCEEDED.
C     N-ORDER OF A.
C     IFAIL-RETURNED 0 IF THE SUBROUTINE HAS EXECUTED CORRECTLY,
C     1 IF NOT.
C     IA-DIMENSION PARAMETER.
C     CAUTION.NEAR-SINGULAR MATRICES MAY GIVE MISLEADING RESULTS.
C     MAXIMUM ORDER OF A=40.
C
C     SUBROUTINE REQUIRED
C             NONE
C
      DOUBLE PRECISION A,BIG,TEST,Q,P
C
      DIMENSION A(IA,IA),P(40),Q(40),IR(40)
C
```

```
      SUBROUTINE EIGS(A,R,EV,N,IA,MV)
C
C     COMPUTES EIGENVALUES AND EIGENVECTORS OF A REAL SYMMETRIC MATRIX
C     USING THRESHOLD JACOBI METHOD.
C     REFERENCE, RALSTON AND WILF, MATHEMATICAL METHODS FOR DIGITAL
C     COMPUTERS, CHAPTER 7.
C     AUTHOR, C.KALLSTROM 1970-07-16.
C
C     A -ORIGINAL MATRIX (SYMMETRIC), DESTROYED IN COMPUTATION.
C        RESULTANT EIGENVALUES ARE DEVELOPED IN DIAGONAL OF MATRIX IN
C        DESCENDING ORDER.
C     R -RESULTANT MATRIX OF EIGENVECTORS (STORED COLUMNWISE, IN SAME
C        SEQUENCE AS EIGENVALUES).
C     EV-VECTOR CONTAINING THE EIGENVALUES IN DESCENDING ORDER.
C     N -ORDER OF MATRICES A AND R.
C     IA-DIMENSION PARAMETER.
C     MV-INPUT CODE
C        0  COMPUTE EIGENVALUES AND EIGENVECTORS.
C        1  COMPUTE EIGENVALUES ONLY (R MUST STILL APPEAR IN CALLING
C           SEQUENCE).
C     THE OFF-DIAGONAL ELEMENTS IN A ARE SET EQUAL TO 0 BEFORE RETURN.
C     THERE ARE NO MAXIMUM ORDER OF THE MATRICES A AND R.
C
C     SUBROUTINE REQUIRED
C            NONE
C
      DIMENSION A(IA,IA),R(IA,IA),EV(1)
C
```

UNIQUENESS OF THE MAXIMUM LIKELIHOOD ESTIMATES OF THE PARAMETERS OF A MIXED AUTOREGRESSIVE MOVING AVERAGE PROCESS

K J ÅSTRÖM
T SÖDERSTRÖM

Report 7306 March 1973
Lund Institute of Technology
Division of Automatic Control

UNIQUENESS OF THE MAXIMUM LIKELIHOOD ESTIMATES OF THE PARAMETERS OF A MIXED AUTOREGRESSIVE MOVING AVERAGE PROCESS.

K.J. Åström and T. Söderström

ABSTRACT.

Estimation of the parameters in a mixed autoregressive moving average process leads to a nonlinear optimization problem. The negative logarithm of the likelihood function, suitably normalized, converges to a deterministic function, called the loss function, as the sample length increases. The local and global extrema of this loss function are investigated. Conditions for the existence of a unique local minimum are given.

TABLE OF CONTENTS. Page

1. INTRODUCTION.

Let $\{y(t), t = 1, 2, \ldots\}$ be a stationary gaussian stochastic process with rational spectral density. It follows from the representation theorem, see e.g. Åström (1970), that the process can be representated as a mixed autoregressive moving average process, i.e.

$$A(q)y(t) = C(q)e(t) \tag{1.1}$$

where e(t) is a sequence of independent normal (0,1) random variables. The operators A(q) and C(q) are given by

$$\begin{cases} A(q) = q^n + a_1q^{n-1} + \ldots + a_n \\ C(q) = q^n + c_1q^{n-1} + \ldots + c_n \end{cases} \tag{1.2}$$

where q is the forward shift operator.

It follows from the representation theorem that the polynomial A(z) can be chosen so that it has all zeros inside the unit circle. The polynomial C(z) may have zeros inside and on the unit circle. The number n can be chosen so that A(q) and C(q) have no common factors.

The estimation of the parameters $a_1, \ldots a_n, c_1, \ldots c_n$ with the maximum likelihood method leads to the problem of minimizing the function

$$V^N(\hat{a}_1, \ldots \hat{a}_n, \hat{c}_1, \ldots \hat{c}_n) = \frac{1}{2N} \sum_{t=1}^{N} \varepsilon^2(t) \tag{1.3}$$

See Åström-Bohlin (1966). The residual $\varepsilon(t)$ is a function of the observations y(1), y(2), ... y(t). It is defined by

$$\varepsilon(t) = \frac{\hat{A}(q)}{\hat{C}(q)}\, y(t) = \frac{\hat{A}(q)C(q)}{\hat{C}(q)A(q)}\, e(t) \tag{1.4}$$

where

$$\begin{cases} \hat{A}(q) = q^n + \hat{a}_1 q^{n-1} + \dots + \hat{a}_n \\ \hat{C}(q) = q^n + \hat{c}_1 q^{n-1} + \dots + \hat{c}_n \end{cases} \tag{1.5}$$

Since $\hat{C}$ and A are assumed to have zeros strictly inside the unit circle and since we only are considering asymptotic properties the initial conditions of (1.4) are not important. They can e.g. be selected as zero.

The maximum likelihood estimates of the model parameters are obtained by finding the absolute minimum of V^N for each N. It can be shown that the estimate will converge to the true parameter values if the polynomials A(z) and $\hat{C}(z)$ have zeros strictly inside the unit circle. Since the function V^N is nonlinear in $\hat{c}_1, \dots, \hat{c}_n$ the minimization must be done numerically. It may happen that the function V^N has several local minima. The existence of local minima may lead to wrong estimates and cause difficulties in the computations.

Since V^N is a random variable it is in general very difficult to analyse the existence of possible local minima. It can, however, be shown that V^N under mild conditions, see Hannan (1960), converges with probability one to the function V, defined by

$$V(\hat{a}_1 \ldots \hat{a}_n, \hat{c}_1 \ldots \hat{c}_n) = \lim_{N\to\infty} V^N(\hat{a}_1 \ldots \hat{a}_n, \hat{c}_1 \ldots \hat{c}_n) =$$

$$= \frac{1}{2} E\varepsilon^2(t) = \frac{1}{4\Pi i} \oint \frac{\hat{A}(z)C(z)\hat{A}(z^{-1})C(z^{-1})}{A(z)\hat{C}(z)A(z^{-1})\hat{C}(z^{-1})} \frac{dz}{z} \qquad (1.6)$$

where the integral path is the unit circle.

The purpose of this report is to find all the local extrema of (1.6).

2. STATEMENT OF THE PROBLEM.

It was previously assumed that

$$\begin{aligned} \circ \quad & n = \deg A(z) = \deg C(z) = \deg \hat{A}(z) = \deg \hat{C}(z) \\ & A(z) \text{ and } C(z) \text{ have no common factors} \end{aligned} \tag{2.1}$$

This condition can be generalized somewhat. For technical reasons it will be suitable to assume that

$$\begin{aligned} \circ \quad & n = \deg A(z) = \deg \hat{A}(z) \\ & m = \deg C(z) = \deg \hat{C}(z) \end{aligned} \tag{2.2}$$

and allow common factors in A and C. It is not necessary that n = m, although this may be a natural choice.

The apparently more general assumption

$$\deg A(z) \leqslant \deg \hat{A}(z)$$

$$\deg C(z) \leqslant \deg \hat{C}(z)$$

is easily obtained from (2.2) putting the last a_i and c_i parameters zero when necessary.

The polynomials involved are now rewritten as

$$\left\{ \begin{aligned} A(z) &= z^n + a_1 z^{n-1} + \ldots + a_n = \prod_1^n (z-\alpha_i) \\ \hat{A}(z) &= z^n + \hat{a}_1 z^{n-1} + \ldots + \hat{a}_n = \prod_1^n (z-\hat{\alpha}_i) \\ C(z) &= z^m + c_1 z^{m-1} + \ldots + c_m = \prod_1^m (z-\gamma_i) \\ \hat{C}(z) &= z^m + \hat{c}_1 z^{m-1} + \ldots + \hat{c}_m = \prod_1^m (z-\hat{\gamma}_i) \end{aligned} \right. \tag{2.3}$$

To establish convergence of V^N it is furthermore assumed that

$$|\alpha_i| < 1 \qquad 1 \leqslant i \leqslant n \tag{2.4}$$

$$|\hat{\gamma}_j| < 1 \qquad 1 \leqslant j \leqslant m \tag{2.5}$$

To ensure that $\hat{a}_i = a_i$, $i = 1, \ldots, n$, $\hat{c}_j = c_j$, $i = 1, \ldots, m$, can be a local minimum the following conditions are assumed

$$|\hat{\alpha}_i| < 1 \qquad 1 \leqslant i \leqslant n \tag{2.6}$$

$$|\gamma_j| < 1 \qquad 1 \leqslant j \leqslant m \tag{2.7}$$

The conditions (2.4) and (2.5) are required to guarantee that the residuals will have finite variance. The condition (2.7) restricts all zeros of C to lie inside the unit circle.

The problem is to find all local extrema of the loss function V (1.6) subject to the constraints (2.4) - (2.7).

3. PRELIMINARIES.

The local extrema of the loss function will now be determined. The calculations are technical but straight-forward. The results are summarized as Lemma 4.1 in Section 4.

Introduce the reciprocials of the polynomials A, $\hat{A}$, C, and $\hat{C}$ defined as

$$\begin{cases} A^*(z) = 1 + a_1 z + \ldots + a_n z^n = z^n A(z^{-1}) \\ \hat{A}^*(z) = 1 + \hat{a}_1 z + \ldots + \hat{a}_n z^n = z^n \hat{A}(z^{-1}) \\ C^*(z) = 1 + c_1 z + \ldots + c_m z^m = z^m C(z^{-1}) \\ \hat{C}^*(z) = 1 + \hat{c}_1 z + \ldots + \hat{c}_m z^m = z^m \hat{C}(z^{-1}) \end{cases} \tag{3.1}$$

The stationary points of V are the solutions of

$$\begin{cases} \frac{\partial V}{\partial \hat{a}_i} = 0 \qquad 1 \leqslant i \leqslant n \\ \\ \frac{\partial V}{\partial \hat{c}_i} = 0 \qquad 1 \leqslant i \leqslant m \end{cases} \tag{3.2}$$

After some computations we find that these conditions can be written as

$$\frac{1}{2\pi i} \oint z^i \frac{\hat{A}(z)C(z)C^*(z)}{A(z)A^*(z)\hat{C}(z)\hat{C}^*(z)} \frac{dz}{z} = 0 \qquad 1 \leqslant i \leqslant n$$
$$\tag{3.3}$$
$$\frac{1}{2\pi i} \oint z^i \frac{\hat{A}(z)\hat{A}^*(z)C(z)C^*(z)}{A(z)A^*(z)\hat{C}(z)\hat{C}^*(z)^2} \frac{dz}{z} = 0 \qquad 1 \leqslant i \leqslant m$$

To avoid the formal difficulty that may arise if A and C have a common factor the polynomials A' and C' are now introduced.

$$\left\{\begin{array}{l} A(z) = A'(z)D(z) \\ C(z) = C'(z)D(z) \\ A'(z) = \prod_{1}^{n-k} (z-\alpha_i) \\ C'(z) = \prod_{1}^{m-k} (z-\gamma_i) \\ D(z) = \prod_{1}^{k}(z-\delta_i) \\ A'(z) \text{ and } C'(z) \text{ relatively prime} \\ \quad (\alpha_i \neq \gamma_j \quad 1 \leqslant i \leqslant n-k,\ 1 \leqslant j \leqslant m-k) \end{array}\right. \tag{3.4}$$

The case $k = 0$ is permitted.

In the same way assume

$$\left\{\begin{array}{l} \hat{A}(z) = \hat{A}'(z)\hat{D}(z) \\ \hat{C}(z) = \hat{C}'(z)\hat{D}(z) \\ \hat{A}'(z) = \prod_{1}^{n-\hat{k}} (z-\hat{\alpha}_i) \\ \hat{C}'(z) = \prod_{1}^{m-\hat{k}} (z-\hat{\gamma}_i) \\ \hat{D}(z) = \prod_{1}^{\hat{k}}(z-\hat{\delta}_i) \\ \hat{A}'(z) \text{ and } \hat{C}'(z) \text{ relatively prime} \end{array}\right. \tag{3.5}$$

Note that the value of $\hat{k}$ depends on the actual point $(\hat{a}_1, \ldots, \hat{a}_n, \hat{c}_1, \ldots, \hat{c}_m)$ in the parameter space.

The polynomials $A'^*(z)$, $\hat{A}'^*(z)$, $C'^*(z)$ and $\hat{C}'^*(z)$ are defined analogous with (3.1).

Furthermore introduce the function

$$f(z) = \frac{\hat{A}'(z)C'(z)C'^*(z)}{A'(z)A'^*(z)\hat{C}'(z)\hat{C}'^*(z)\hat{C}^*(z)} \qquad (3.6)$$

Using (3.4) - (3.6) the equations (3.3) are rewritten

$$\begin{cases} \dfrac{1}{2\pi i}\displaystyle\oint z^i\hat{C}'^*(z)f(z)\frac{dz}{z} = 0 & 1 \leqslant i \leqslant n \\ \dfrac{1}{2\pi i}\displaystyle\oint z^i\hat{A}'^*(z)f(z)\frac{dz}{z} = 0 & 1 \leqslant i \leqslant m \end{cases} \qquad (3.7)$$

The definition of $\hat{A}'^*(z)$ and $\hat{C}'^*(z)$ gives

$$\begin{cases} \displaystyle\sum_{j=0}^{m-\hat{k}} \hat{c}_j' \frac{1}{2\pi i}\oint z^{i+j}f(z)\frac{dz}{z} = 0 & 1 \leqslant i \leqslant n \\ \displaystyle\sum_{j=0}^{n-\hat{k}} \hat{a}_j' \frac{1}{2\pi i}\oint z^{i+j}f(z)\frac{dz}{z} = 0 & 1 \leqslant i \leqslant m \end{cases} \qquad (3.8)$$

Define for $p \geqslant 1$

$$F_p = \frac{1}{2\pi i}\oint z^p f(z)\frac{dz}{z} \qquad (3.9)$$

Then (3.8) becomes

$$\left[\begin{array}{cccccccc}
1 & \hat{c}'_1 & \cdots & \hat{c}'_{m-\hat{k}} & & & 0 & \\
 & \ddots & & & \ddots & & & \\
0 & & \ddots & & & \ddots & & \\
 & & & 1 & \hat{c}'_1 & & & \hat{c}'_{m-\hat{k}} \\
\hline
1 & \hat{a}'_1 & \cdots & \cdots & \cdots & \hat{a}'_{n-\hat{k}} & & \\
 & \ddots & & & & & 0 & \\
0 & & \ddots & & & & \ddots & \\
 & & 1 & \hat{a}'_1 & \cdots & \cdots & & \hat{a}'_{n-\hat{k}}
\end{array}\right]
\left[\begin{array}{c}
F_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ F_{n+m-\hat{k}}
\end{array}\right] = 0 \qquad (3.10)$$

The matrix in (3.10) is $(n+m) \times (n+m-\hat{k})$. Since $\hat{A}'(z)$ and $\hat{C}'(z)$ by assumption are relatively prime it follows from elementary algebra that the rank of the matrix is $n+m-\hat{k}$. See e.g. Dickson (1922).

Thus

$$\frac{1}{2\Pi i} \oint z^{i} f(z) \frac{dz}{z} = 0 \qquad 1 \leqslant i \leqslant n+m-\hat{k} \qquad (3.11)$$

The poles of $f(z)$ inside the unit circle now are relabelled through

$$A'(z)\hat{C}'(z) = \prod_{1}^{n-k} (z-\alpha_i) \prod_{1}^{m-\hat{k}} (z-\hat{\gamma}_j) = \prod_{1}^{\ell} (z-u_i)^{t_i} \qquad (3.12)$$

where $u_i \neq u_j$ if $i \neq j$, $t_i \geqslant 1$ all i and

$$\sum_{1}^{\ell} t_i = n + m - k - \hat{k} \qquad (3.13)$$

This implies that $f(z)$ can be written

$$f(z) = \frac{g(z)}{\prod_{i=1}^{\ell} (z-u_i)^{t_i}} \qquad (3.14)$$

where

$$g(z) = \hat{A}'(z)C'(z) \cdot \frac{C'^{*}(z)}{A'^{*}(z)\hat{C}'^{*}(z)\hat{C}^{*}(z)} \tag{3.15}$$

is analytic inside the unit circle.

Using (3.14), the equation (3.11) can be replaced by

$$0 = \frac{1}{2\Pi i} \oint \frac{z^{i-1} g(z)}{\prod\limits_{1}^{\ell} (z-u_j)^{t_j}} dz = \sum_{k=1}^{\ell} \operatorname*{Res}_{z=u_k} \frac{z^{i-1} g(z)}{\prod\limits_{j=1}^{\ell} (z-u_j)^{t_j}} =$$

$$= \sum_{k=1}^{\ell} \frac{1}{(t_k-1)!} D^{(t_k-1)} \left[\frac{z^{i-1} g(z)}{\prod\limits_{\substack{j=1 \\ j \neq k}}^{\ell} (z-u_j)^{t_j}} \right]_{z=u_k} =$$

$$= \sum_{k=1}^{\ell} \frac{1}{(t_k-1)!} \sum_{\nu=0}^{t_k-1} \binom{t_k-1}{\nu} D^{(\nu)} [z^{i-1}]_{z=u_k} D^{(t_i-1-\nu)} \left[\frac{g(z)}{\prod\limits_{j \neq k} (z-u_j)^{t_j}} \right]_{z=u_k}$$

where D denotes differentiation with respect to z.

Hence

$$\sum_{k=1}^{\ell} \sum_{\nu=0}^{t_k-1} D^{(\nu)} [z^{i-1}]_{z=u_k} \cdot d_{k\nu} = 0 \qquad 1 \leq i \leq n+m-\hat{k} \tag{3.16}$$

where

$$d_{k\nu} = \frac{1}{\nu!(t_k-1-\nu)!} D^{(t_k-1-\nu)} \left[\frac{g(z)}{\prod\limits_{j \neq k} (z-u_j)^{t_k}} \right]_{z=u_k} \tag{3.17}$$

Using matrix notation (3.16) can be expressed as

$$S \cdot G = 0 \tag{3.18}$$

where G is a $(m+n-k-\hat{k})$ vector,

$$G = \begin{bmatrix} d_{10} \\ \vdots \\ d_{1,t_1-1} \\ d_{20} \\ \cdot \\ \cdot \\ \cdot \\ d_{\ell,t_\ell-1} \end{bmatrix}$$

and S a $(m+n-\hat{k}) \times (m+n-k-\hat{k})$ matrix

$$S = \begin{bmatrix} 1 & 0 & \cdots & 1 & \cdots & D^{(t_\ell-1)}[z^0]_{z=u_\ell} \\ u_1 & 1 & & u_2 & & D^{(t_\ell-1)}[z^1]_{z=u_\ell} \\ u_1^2 & 2u_1 & & u_2^2 & & D^{(t_\ell-1)}[z^2]_{z=u_\ell} \\ \cdot & \cdot & & \cdot & & \\ \cdot & \cdot & & \cdot & & \\ \cdot & \cdot & & \cdot & & \\ u_1^{m+n-\hat{k}-1} & (m+n-\hat{k}-1)u_1^{m+n-\hat{k}-2} & \cdots & u_2^{m+n-\hat{k}-1} & \cdots & D^{(t_\ell-1)}[z^{m+n-\hat{k}-1}]_{z=u_\ell} \end{bmatrix}$$

The matrix S is a generalization of the van der Monde matrix. It follows from Kaufman (1969) that its upper, square part is non-singular. Thus (3.18) implies

$$G = 0 \tag{3.19}$$

A useful lemma will now be proved.

Lemma 3.1. Let h(z) and g(z) be analytic functions in a neighbourhood of z_0. Assume that $g(z_0) \neq 0$. Then

$$D^{(p)}[h(z)g(z)]_{z=z_0} = 0 \qquad 0 \leqslant p \leqslant n \tag{3.20}$$

is equivalent to

$$D^{(p)}[h(z)]_{z=z_0} = 0 \qquad 0 \leqslant p \leqslant n \tag{3.21}$$

Proof. The equation (3.20) implies

$$\sum_{i=0}^{p} \binom{p}{i} D^{(i)}[h(z)]_{z=z_0} D^{(p-i)}[g(z)]_{z=z_0} = 0 \qquad 0 \leqslant p \leqslant n$$

or equivalent in matrix form

$$\begin{bmatrix} D^{(0)}g(z) & & & & 0 \\ D^{(1)}g(z) & D^{(0)}g(z) & & & \\ D^{(2)}g(z) & 2D^{(1)}g(z) & D^{(0)}g(z) & & \\ \vdots & & & \ddots & \\ D^{(n)}g(z) & \cdot & \cdot & \cdots & D^{(0)}g(z) \end{bmatrix}_{z=z_0} \cdot$$

$$\cdot \begin{bmatrix} D^{(0)}h(z) \\ D^{(1)}h(z) \\ \vdots \\ D^{(n)}h(z) \end{bmatrix}_{z=z_0} = 0 \tag{3.22}$$

According to the assumption $g(z_0) \neq 0$ the matrix is non-singular and the equivalence between (3.20) and (3.21) follows.

□

Equations (3.19) and (3.17) give

$$D^{(t_k-1-\nu)}\left[\frac{g(z)}{\prod_{j\neq k}(z-u_j)^{t_k}}\right]_{z=u_k} = 0 \tag{3.23}$$

$$0 \leqslant \nu \leqslant t_k-1,\ 1 \leqslant k \leqslant \ell$$

and it follows from Lemma 3.1 that

$$\begin{cases} D^{(i)}[g(z)]_{z=u_k} = 0 \\ 0 \leqslant i \leqslant t_{k-1},\ 1 \leqslant k \leqslant \ell \end{cases} \tag{3.24}$$

Using the Lemma 3.1 again [with $h(z) = \hat{A}'(z)C'(z)$ cf. (3.15)] the following equations are obtained.

$$\begin{cases} D^{(i)}[\hat{A}'(z)C'(z)]_{z=u_k} = 0 \\ 0 \leqslant i \leqslant t_{k-1},\ 1 \leqslant k \leqslant \ell \end{cases} \tag{3.25}$$

Hence

$$\hat{A}'(z)C'(z) \equiv \prod_{k=1}^{\ell}(z-u_k)^{t_k-1} \equiv A'(z)\hat{C}'(z) \tag{3.26}$$

Thus it has been shown that the stationary points, i.e. the solutions of (3.7) must fulfil (3.26). Conversely, the calculations show that (3.26) implies (3.7). The latter assertion can be proven directly since (3.26) implies that f(z) has no poles inside the unit circle.

4. MAIN RESULT.

The following lemma is a summary of the calculations in the previous section.

Lemma 4.1. Consider the loss function (1.6) subject to the conditions (2.4), (2.7) and the constraints (2.5), (2.6). Let $A'(z)$, $C'(z)$, $\hat{A}'(z)$, $\hat{C}'(z)$ be defined by (3.4), (3.5). Then the stationary points of V are the solutions of

$$\hat{A}'(z)C'(z) \equiv A'(z)\hat{C}'(z) \qquad (4.1)$$

The next lemma deals with global minimum points.

Lemma 4.2. Consider the loss function V (1.6) subject to the conditions (2.4), (2.7) and the constraints (2.5), (2.6). Then the global minimum points of V are the solutions of

$$\hat{A}'(z)C'(z) \equiv A'(z)\hat{C}'(z) \qquad (4.1)$$

Proof. Introduce

$$H^*(z) = \frac{\hat{A}^*(z)C^*(z)}{A^*(z)\hat{C}^*(z)} = 1 + \sum_{i=1}^{\infty} h_i z^i$$

where the infinite series converges in and on the unit circle. Put $h_0 = 1$, then

$$V = \frac{1}{4\pi i} \oint \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} h_j z^j h_k z^{-k} \frac{dz}{z} = \frac{1}{2}\left(1 + \sum_{j=1}^{\infty} h_j^2\right)$$

Thus

$V \geq 1/2$

with equality if and only if $H(z) \equiv 1$ or

$\hat{A}(z)C(z) \equiv A(z)\hat{C}(z)$

Invoking (3.4) and (3.5) we find that this equation is equivalent to (4.1).

□

It remains to analyze the solution of (4.1). The equation can be written as

$$\frac{C'(z)}{A'(z)} \equiv \frac{\hat{C}'(z)}{\hat{A}'(z)} \tag{4.2}$$

The number $\hat{k}$ has not been determined yet. To establish equality in (4.2) for all z it is necessary that both sides have the same poles and zeros. Since there are no common factor it is thus necessary and sufficient that $\hat{k} = k$, $\hat{A}'(z) \equiv A(z)$ and $\hat{C}'(z) = C(z)$.

Two cases can be separated:

1. $k = 0$. Then $\hat{k} = 0$ and the loss function has a unique local minimum

2. $k > 0$. Then $\hat{k} > 0$ and there are infinite many local minimum points. In fact, these minimum points form a manifold in the parameter space. On this manifold the loss function obtains its infimum. This case means that the model contains too many parameters.

Another way of characterization is the following. The unknown parameters are $\hat{k}$, $\hat{a}'_1$, ..., $\hat{a}'_{n-\hat{k}}$, $\hat{c}'_1$, ..., $\hat{c}'_{m-\hat{k}}$, $\hat{d}_1$, ..., $\hat{d}_{\hat{k}}$. Of these must for all minimum points

$$\begin{cases} \hat{k} = k \\ \hat{a}_i' = a_i' & 1 \leq i \leq n-\hat{k} \\ \hat{c}_i' = c_i' & 1 \leq i \leq m-\hat{k} \end{cases}$$

while $\hat{d}_1 \ldots \hat{d}_{\hat{k}}$ (if $\hat{k} > 0$) are arbitrary.

The result of the calculation and the discussion is summed up in the following theorem.

Theorem. Consider the loss function (1.6) subject to the conditions (2.4), (2.7) and the constraints (2.5) and (2.6). Assume that deg $A(z) = n$, deg $\hat{A}(z) = n+\hat{n}$, deg $C(z) = m$, deg $\hat{C}(z) = m+\hat{m}$ where $\min(\hat{n},\hat{m}) \geq 0$ and that $A(z)$ and $C(z)$ are relatively prime.

i) If $\min(\hat{n},\hat{m}) = 0$ there is a unique local minimum, namely

$$\hat{a}_i = \begin{cases} a_i & 1 \leq i \leq n \\ 0 & \text{if } i > n \text{ and } \hat{n} > 0 \end{cases}$$

$$\hat{c}_i = \begin{cases} c_i & 1 \leq i \leq m \\ 0 & \text{if } i > m \text{ and } \hat{m} > 0 \end{cases}$$

ii) If $\min(\hat{n},\hat{m}) > 0$ there are infinitely many local minimum points given by the manifold

$$\left.\begin{aligned} \hat{A}(z) &\equiv L(z)A(z) \\ \hat{C}(z) &\equiv L(z)C(z)z^{\hat{m}-\hat{n}} \end{aligned}\right\} \quad \text{if } \hat{m} \geq \hat{n}$$

or

$$\left.\begin{aligned}\hat{A}(z) &\equiv L(z)A(z)z^{\hat{n}-\hat{m}} \\ \hat{C}(z) &\equiv L(z)C(z)\end{aligned}\right\} \quad \text{if } \hat{m} \leq \hat{n}$$

L(z) is an arbitrary unitary polynomial of degree $\min(\hat{n},\hat{m})$. Each point in the manifold also is a global minimum point.

iii) There are neither local maxima nor saddle points.

5. ACKNOWLEDGEMENTS.

The authors want to thank Mr. I. Gustavsson for many stimulating discussions.

It is also a pleasure to thank Mrs. G. Christensen who typed the manuscript. The partial support of the Swedish Board for Technical Development under contract 72-202/U137 is gratefully announced.

6. REFERENCES.

Åström, K.J. (1970).
Introduction to Stochastic Control Theory. Academic Press.

Åström, K.J. - Bohlin, T. (1966).
Numerical Identification of Linear Dynamical Systems from Normal Operating Records. Paper, IFAC Symposium on Theory of Self-Adaptive Systems, Teddington, England. In Theory of Self-Adaptive Control Systems (Ed. P.H. Hammond), Plenum Press, New York.

Dickson, L.E. (1922).
First Course in the Theory of Equations. Wiley.

Hannan, E.J. (1960).
Time Series Analysis. Meuthen and Co., London.

Kaufman, I. (1969).
The Inversion of the Vandermonde Matrix and the Transformation to the Jordan Canonical Form. IEEE Trans. Aut. Control, AC-14, p. 774-777.

ON THE UNIQUENESS OF MAXIMUM LIKELIHOOD IDENTIFICATION FOR DIFFERENT STRUCTURES

T SÖDERSTRÖM

Report 7307 March 1973
Lund Institute of Technology
Division of Automatic Control

ON THE UNIQUENESS OF MAXIMUM LIKELIHOOD IDENTIFICATION FOR DIFFERENT STRUCTURES.

T. Söderström

ABSTRACT.

Maximum likelihood identification of a linear dynamic system is performed as a minimization of a loss function. The concept of uniqueness of the parameter estimates is closely related to the number of local minimum points of this loss function. The number of local minimum points is examined for some different models. Asymptotic expressions for the loss function are used. Conditions are given which imply a unique local minimum point.

TABLE OF CONTENTS

I. INTRODUCTION.

The maximum likelihood (ML) method is a useful tool for estimation of parameters in system equations. The ML estimate $\hat{\theta}_{ML}$ is the global maximum point of the likelihood function $L(\hat{\theta})$, i.e.

$$L(\hat{\theta}_{ML}) \geq L(\hat{\theta}) \text{ all } \hat{\theta}$$

In most cases there is no analytical expression for the maximum point $\hat{\theta}_{ML}$. The maximization of $L(\hat{\theta})$ has to be done computationally using some search routine. Such a search routine may converge to a local maximum point θ^* of $L(\hat{\theta})$, i.e.

$$L(\theta^*) \geq L(\hat{\theta}) \quad \text{all } \hat{\theta} \text{ close to } \theta^*$$

It is then valuable to know if the likelihood function has a unique local maximum point or not.

This issue is closely related to the concept of identifiability, see Bellman-Åström (1970). The purpose of this report is to analyze the local maximum points of the likelihood function for some different structures. Bohlin (1971) has given some tests, which can be used for detecting if a local maximum or generally an arbitrary point is a global maximum point or not.

The report is organized as follows: In this chapter some basic assumptions are given. In the next chapter the mathematical tools of the analysis are penetrated. Chapter III contains an examination of the global maximum points for the different structures. It is desirable that the true value θ is a global maximum point and preferably a unique one.

Moreover, this examination simplifies the analysis of the local maximum points, since it describes all "desirable" points. The last three chapters deal with the examination of the local maximum points for some specific likelihood functions.

Consider a system given by

$$y(t) = G(\theta;q^{-1})u(t) + H(\theta;q^{-1})e(t)$$

where

$$G(\theta;q^{-1}) = \sum_{0}^{\infty} g_i(\theta)q^{-i}$$

$$H(\theta;q^{-1}) = \sum_{0}^{\infty} h_i(\theta)q^{-i}$$

u(t) is the input, y(t) the output and e(t) gaussian white noise with zero mean and standard deviation λ. q^{-1} is the backward shift operator. It is assumed that $h_0(\theta) \equiv 1$. The system can be illustrated by the figure below.

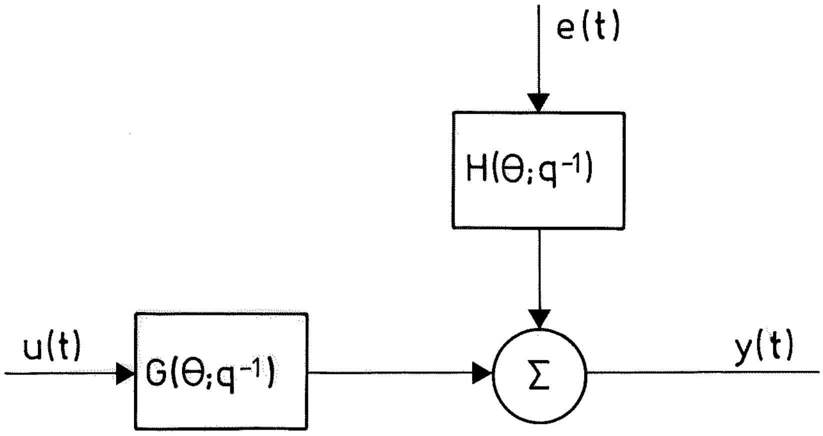


Figure 1 - Block diagram of the system.

The purpose of an identification is to estimate the value of the vector θ based on an input-output record. The true value will be denoted by θ.

In this report some different transfer functions G and H will be considered. It is assumed that G and H are rational functions in q^{-1}. The coefficients are functions of θ.

Under these assumptions the maximization of the likelihood function is equivalent to the minimization of the loss function, see Åström-Bohlin (1966).

$$V(\hat{\theta},\theta) = \frac{1}{2N} \sum_{t=1}^{N} \varepsilon^2(t) \tag{1.1}$$

where the residuals $\varepsilon(t)$ are defined by

$$y(t) = G(\hat{\theta};q^{-1})u(t) + H(\hat{\theta};q^{-1})\varepsilon(t) \tag{1.2}$$

while the output is given from

$$y(t) = G(\theta;q^{-1})u(t) + H(\theta;q^{-1})e(t) \tag{1.3}$$

The ML estimate $\hat{\theta}_{ML}$ of θ is thus given by

$$V(\hat{\theta}_{ML},\theta) = \min_{\hat{\theta}} V(\hat{\theta},\theta)$$

assuming that a global minimum exists.

The residuals can be written as

$$\varepsilon(t) = \frac{G(\theta;q^{-1}) - G(\hat{\theta};q^{-1})}{H(\hat{\theta};q^{-1})} u(t) + \frac{H(\theta;q^{-1})}{H(\hat{\theta};q^{-1})} e(t) \tag{1.4}$$

In the analysis of the loss function (1.1) ergodic theory will be used.

The generalized least squares method has been treated elsewhere by the author in Söderström (1972), where it is shown that the loss function in this case has a unique local minimum point when the signal to noise ratio is high enough. For small values of this ratio there may exist several local minimum points.

For the other cases treated here it is shown (under suitable assumptions) that all local minimum points are global minimum points. There will be a unique global (and local) minimum point if the correct order of the transfer functions is used.

II. MATHEMATICAL PRELIMINARIES.

In this chapter the basic mathematical tools for the analysis of the loss functions are given.

First some conventions used in the report are presented. Then some polynomial equations are studied. A lemma giving sufficient conditions for the existence of

$$\lim_{N\to\infty} V(\hat{\theta},\theta)$$

is considered. Finally the concept of persistently exciting signals is treated and some applications are made. Some of the lemmas are given in Söderström (1972). They are stated here too in order to clarify their use in the analysis.

In order to simplify the notations the following conventions will be used throughout the report.

Convention 2.1. Polynomial operators will be denoted by capital letters, e.g. $A(q^{-1})$. The number of coefficients will be denoted by n or $\hat{n}$ with a corresponding lower case letter as a subscript.

Examples:

$$A(q^{-1}) = 1 + \sum_{1}^{n_a} a_i q^{-i} \qquad \hat{A}(q^{-1}) = 1 + \sum_{1}^{\hat{n}_a} \hat{a}_i q^{-i}$$

$$B(q^{-1}) = \sum_{1}^{n_b} b_i q^{-i} \qquad \hat{B}(q^{-1}) = \sum_{1}^{\hat{n}_b} \hat{b}_i q^{-i}$$

The expression

$$\sum_{1}^{n_a} a_i q^{-i}$$

is interpreted as zero if $n_a = 0$.

Convention 2.2. Given two polynomials

$$A(z) = \sum_{i=0}^{n_a} a_i z^i \qquad\qquad B(z) = \sum_{i=0}^{n_b} b_i z^i$$

the notation $A(z) \equiv B(z)$ means

$$a_i = b_i \qquad 0 \leqslant i \leqslant \min(n_a, n_b)$$

and

$$\begin{array}{lll} \text{if } n_a > n_b & a_i = 0 & n_b < i \leqslant n_a \\ \text{if } n_b > n_a & b_i = 0 & n_a < i \leqslant n_a \end{array}$$

Convention 2.3. Given the polynomials A(z) and B(z) (and C(z)). They are said to be relatively prime if there is no common factor to all the polynomials. The physical interpretation is that the system

$$A(q^{-1})y(t) = B(q^{-1})u(t)$$

$$\left(A(q^{-1})y(t) = B(q^{-1})u(t) + C(q^{-1})e(t)\right)$$

is controllable and observable.

Convention 2.4. $\mathcal{E}x(t)$ denotes

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} x(t)$$

If x(t) is an ergodic stochastic process $\mathcal{E}x(t) = Ex(t)$. All stochastic processes in this report are ergodic. The notation is used for deterministic signals as well.

The following elementary two lemmas from the theory of equations will be useful. The proofs are not very difficult and they are given here.

The first lemma deals with an equation, which will occur several times in the forthcoming analysis.

Lemma 2.1. Given the polynomials

$$A(z) = 1 + \sum_{i=1}^{n_a} a_i z^i$$

and

$$B(z) = \sum_{i=1}^{n_b} b_i z^i$$

Consider the following equation in the unknowns $(\hat{a}_1, \ldots, \hat{a}_{\hat{n}_a}, \hat{b}_1, \ldots, \hat{b}_{\hat{n}_b})$ with $n_\ell = \min(\hat{n}_a - n_a, \hat{n}_b - n_b) \geq 0$

$$\hat{A}(z)B(z) - A(z)\hat{B}(z) \equiv 0 \qquad (2.1)$$

Assume that A(z) and B(z) are relatively prime.

i) If $n_\ell = 0$ the only solution is given by

$$\hat{A}(z) \equiv A(z)$$
$$\hat{B}(z) \equiv B(z) \tag{2.2}$$

ii) If $n_\ell > 0$ all solutions are given by

$$\hat{A}(z) \equiv A(z)L(z)$$
$$\hat{B}(z) \equiv B(z)L(z) \tag{2.3}$$

where

$$L(z) = 1 + \sum_{i=1}^{n_\ell} \ell_i z^i$$

The coefficients $(\ell_i)_1^{n_\ell}$ are arbitrary.

Proof. Since $A(z) \not\equiv 0$, $\hat{A}(z) \not\equiv 0$ the equation can be written

$$\frac{B(z)}{A(z)} = \frac{\hat{B}(z)}{\hat{A}(z)} \quad \text{all } z$$

Noting that the right hand side must have the same zeros and poles as the left hand side the assertions are obvious.

Q.E.D.

Corr. If $B(z)$ is of the form

$$B(z) = 1 + \sum_{1}^{n_b} b_i z^i$$

and $\hat{B}(z)$ of the form

$$\hat{B}(z) = 1 + \sum_{1}^{\hat{n}_b} \hat{b}_i z^i$$

the lemma remains true without changes.

Lemma 2.2. Consider the following matrix of order $\max(\hat{n}_a+n_b,\ n_a+\hat{n}_b) \times (\hat{n}_a+\hat{n}_b)$

$$P = \left[\begin{array}{cccc|cccc}
0 & & & & 1 & & & \\
b_1 & & 0 & & \cdot & \ddots & 0 & \\
\vdots & \ddots & & & \vdots & & \ddots & 1 \\
b_{n_b} & & \ddots & & a_{n_a} & & & \vdots \\
 & \ddots & & b_1 & & \ddots & & \vdots \\
 & 0 & \ddots & \vdots & & 0 & \ddots & \vdots \\
 & & & b_{n_b} & & & & a_{n_a}
\end{array}\right] \tag{2.4}$$

$\hat{n}_a$ columns $\qquad\qquad$ $\hat{n}_b$ columns

(At least one of the figures b_{n_b}, a_{n_a} is on the last row.)

Let $A(z)$ and $B(z)$ have m common zeros. Assume that $\hat{n}_a \geq n_a$, $\hat{n}_b \geq n_b$.
Then rank $P = \max(\hat{n}_a+n_b,\ n_a+\hat{n}_b) - m$.

Proof. Consider the equation

$$\hat{A}(z)B(z) - A(z)\hat{B}(z) \equiv 0$$

From lemma 2.1 it is known that the general solution is of the form

$$\hat{A}(z) \equiv \bar{A}(z)L(z)$$

$$\hat{B}(z) \equiv \bar{B}(z)L(z)$$

where

$\bar{A}(z)$ and $\bar{B}(z)$ are relatively prime

$$L(z) = 1 + \sum_{1}^{n_\ell} \ell_i z^i$$

$$n_\ell = \min(\hat{n}_a - n_a, \hat{n}_b - n_b) + m$$

Introduce new variables $c_1 \ldots c_{\hat{n}_a}$, $d_1 \ldots d_{\hat{n}_b}$ by

$$C(z) = \sum_{1}^{\hat{n}_a} c_i z^i \equiv \hat{A}(z) - \bar{A}(z)$$

$$D(z) = \sum_{1}^{\hat{n}_b} d_i z^i \equiv \hat{B}(z) - \bar{B}(z)$$

The equation is then

$$C(z)B(z) - A(z)D(z) \equiv 0$$

with the general solution

$$C(z) \equiv \bar{A}(z)\big(L(z) - 1\big)$$

$$D(z) = \bar{B}(z)\big(L(z) - 1\big)$$

However, this equation can be written as

$$P \cdot \begin{bmatrix} c_1 \\ \vdots \\ c_{\hat{n}_a} \\ -d_1 \\ \vdots \\ -d_{\hat{n}_b} \end{bmatrix} = 0$$

The expression of the general solution implies that $\dim N(P) = n_\ell$.
Thus rank P is given by

$$= \dim R(P^T) = \hat{n}_a + \hat{n}_b - \dim N(P)$$

$$= \hat{n}_a + \hat{n}_b - \min(\hat{n}_a - n_a,\ \hat{n}_b - n_b) - m$$

$$= \max(\hat{n}_a + n_b,\ n_a + \hat{n}_b) - m$$

Q.E.D.

Remark. In the case $\hat{n}_a = n_a$, $\hat{n}_b = n_b$ (P is square) P is nonsingular if and only if $m = 0$. This fact is already shown by e.g. Dickson (1922). In this report, however, the general case will be needed.

In the analysis of the loss functions ergodic expressions will be used. The loss functions are all of the form

$$\frac{1}{2N} \sum_{t=1}^{N} \varepsilon^2(t)$$

with $\varepsilon(t)$ given by (1.4). The following lemma gives sufficient conditions for convergence of such expressions.

Lemma 2.3. Consider the system

$$y(t) = G(q^{-1})u(t) + H(q^{-1})e(t)$$

where $G(q^{-1})$ and $H(q^{-1})$ are asymptotically stable filters of finite orders, and $e(t)$ is white noise with finite fourth moment, independent of $u(t)$.

The input $u(t)$ is the sum of two terms, $u_1(t)$ and $u_2(t)$, of which one may vanish. The term $u_1(t)$ is deterministic such that to every $\varepsilon > 0$ there is a periodic function $u_1'(t)$ fulfilling

$$|u_1(t) - u_1'(t)| < \varepsilon \quad \text{all } t$$

The second term is given by

$$u_2(t) = F(q^{-1})v(t)$$

where $F(q^{-1})$ is an asymptotically stable filter of finite order and $v(t)$ white noise with finite fourth moment.

Let $D_1(q^{-1})$ and $D_2(q^{-1})$ be two arbitrary asymptotically stable filters of finite order. Then

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} [D_1(q^{-1})y(t) + D_2(q^{-1})u(t)] \begin{bmatrix} y(t) \\ u(t) \end{bmatrix} \qquad (2.5)$$

exists with probability one and in mean square.

If $u(t)$ and $y(t)$ are stochastic processes the limit is

$$E[D_1(q^{-1})y(t) + D_2(q^{-1})u(t)] \begin{bmatrix} y(t) \\ u(t) \end{bmatrix}$$

Proof. See Söderström (1972).

The notion of persistent excitation introduced in Åström--Bohlin (1966) is very useful in the analysis of the loss function.

Definition 2.1. u(t) is said to be persistently exciting of order n if

i) $$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} u(t) = \bar{u} \quad \text{and}$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} [u(t) - \bar{u}][u(t+\tau) - \bar{u}] = r_u(\tau)$$

exist and

ii) the n by n symmetric matrix

$$R_u = \begin{bmatrix} r_u(0) & r_u(1) & \dots & r_u(n-1) \\ & \ddots & & \vdots \\ & & \ddots & \vdots \\ & & & r_u(0) \end{bmatrix}$$

is positive definite.

Some simple properties of persistently exciting signals and a characterization of this concept in the frequency domain is given in Ljung (1971). In this report the following properties will be used (proved in Ljung (1971)).

Lemma 2.4. u(t) is persistently exciting of order n if and only if the spectral density corresponding to the sample covariance function is non zero (in distributive sense) in at least n different points.

If $u(t)$ is periodic, the spectral density will be discrete and consist of a number of δ-functions. The distribution $\delta(x)$ is here considered as non zero in $x = 0$.

Corr. Let $y(t) = H(q^{-1})u(t)$. If $u(t)$ is persistently exciting of order n and $H(q^{-1})$ is stable and has no zeros on the unit circle, then $y(t)$ is persistently exciting of order n.

A simple application is made in

Lemma 2.5. Let

$$y(t) = H(q^{-1})u(t)$$

$$H(q^{-1}) = \sum_{i=0}^{n-1} h_i q^{-i}$$

i) If $y(t) \equiv 0$ with probability one and $u(t)$ is persistently exciting of order n, then $h_i = 0$, $i = 0,\dots,n-1$.

ii) If $u(t)$ is not persistently exciting of order n, then there exists $H(q^{-1}) \not\equiv 0$ such that $y(t) \equiv 0$ with probability one.

Proof. See Söderström (1972).

A combination of Lemma 2.1 and Lemma 2.5 gives a further result.

Lemma 2.6. Given $A(q^{-1})$, $B(q^{-1})$ and $u(t)$. Assume that $A(q^{-1})$ and $B(q^{-1})$ are relatively prime and that $n_\ell = \min(\hat{n}_a - n_a,\ \hat{n}_b - n_b) \geq 0$.
Consider the equation:

$$[\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1})]u(t) = 0 \quad \text{a.s.} \tag{2.6}$$

Let $m = \max(\hat{n}_a + n_b,\ n_a + \hat{n}_b)$.

i) If $u(t)$ is persistently exciting of order m the general solution is given by

$$\begin{cases} \hat{A}(q^{-1}) \equiv A(q^{-1})L(q^{-1}) \\ \hat{B}(q^{-1}) \equiv B(q^{-1})L(q^{-1}) \end{cases} \tag{2.7}$$

where

$$L(q^{-1}) = \begin{matrix} 1 + \sum_{1}^{n_\ell} \ell_i q^{-1} & \text{if } n_\ell \geq 1 \\ 1 & \text{if } n_\ell = 0 \end{matrix}$$

The numbers ℓ_i are arbitrary.

ii) If $u(t)$ is not persistently exciting of order m there is at least one more solution of (2.6) than (2.7).

Proof. If $u(t)$ is persistently exciting of order m it follows from Lemma 2.5 that

$$\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1}) \equiv 0$$

The general solution is then obtained from Lemma 2.1.

If $u(t)$ is not persistently exciting of order m, Lemma 2.5 implies the existence of

$$H(q^{-1}) = \sum_{1}^{m} h_i q^{-i} \not\equiv 0$$

such that

$$H(q^{-1})u(t) = 0$$

Writing the equation

$$\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1}) \equiv H(q^{-1})$$

and invoking Lemma 2.2 the assertion ii) follows.

Q.E.D.

The concept of persistent excitation is now applied to a matrix consisting of covariances of the input and the output.

Definition 2.1. Let $y(t) = G(q^{-1})u(t)$. The following matrix of order $(m_a+m_b) \times (m_a+m_b)$ will be called the system covariance matrix of type (m_a, m_b).

$$R = \left[\begin{array}{ccc|ccc}
r_y(0) & \cdot\ \cdot & r_y(m_a-1) & -r_{yu}(0) & \cdot\ \cdot & -r_{yu}(m_b-1) \\
\vdots & & \vdots & \vdots & & \vdots \\
r_y(m_a-1) & \cdot\ \cdot & r_y(0) & -r_{yu}(1-m_a) & \cdot\ \cdot & -r_{yu}(m_b-m_a) \\
\hline
-r_{yu}(0) & \cdot\ \cdot & -r_{yu}(1-m_a) & r_u(0) & \cdot\ \cdot\ \cdot & r_u(m_b-1) \\
\vdots & & \vdots & \vdots & & \vdots \\
-r_{yu}(m_b-1) & \cdot\ \cdot & -r_{yu}(m_b-m_a) & r_u(m_b-1) & \cdot\ \cdot & r_u(0)
\end{array}\right]$$

Lemma 2.7. Let

$$y(t) = \frac{B(q^{-1})}{A(q^{-1})} u(t)$$

where $A(q^{-1})$ and $B(q^{-1})$ are relatively prime. Consider the system covariance matrix R of type (m_a, m_b). Assume that u(t) is persistently exciting of order $\max(m_a+n_b, n_a+m_b)$ and let $n_\ell = \min(m_a-n_a, m_b-n_b)$.

i) Then R is positive definite if and only if $n_\ell \leqslant 0$.

ii) If $n_\ell > 0$ the null space of R has dimension n_ℓ and is spanned by vectors of the following form:

$[c_1 \ldots c_{m_a}, d_1 \ldots d_{m_b}]^T$ with

$$C(q^{-1}) = \sum_1^{m_a} c_i q^{-i} \equiv A(q^{-1})L(q^{-1})$$

$$D(q^{-1}) = \sum_1^{m_b} d_i q^{-i} \equiv B(q^{-1})L(q^{-1})$$

$$L(q^{-1}) = \sum_1^{n_\ell} \ell_i q^{-i}$$

The numbers ℓ_i are arbitrary.

Proof. In order to investigate the null space of R consider the equation

$$x^T R x = 0 \qquad (2.8)$$

Let $x^T = [c_1 \; \dots \; c_{m_a} \; d_1 \; \dots \; d_{m_b}]$ and introduce the corresponding operators

$$C(q^{-1}) = \sum_1^{m_a} c_i q^{-i}, \quad D(q^{-1}) = \sum_1^{m_b} d_i q^{-i}$$

Then

$$x^T R x = E\Big[[-y(t-1) \; \dots \; -y(t-m_a) u(t-1) \; \dots \; u(t-m_b)]x\Big]^2$$

$$= E\big[-C(q^{-1})y(t) + D(q^{-1})u(t)\big]^2$$

The equation (2.8) is thus equivalent to

$$\frac{C(q^{-1})B(q^{-1}) - D(q^{-1})A(q^{-1})}{A(q^{-1})} u(t) = 0 \text{ a.s.}$$

From Lemma 2.4 Corr and Lemma 2.6 it follows that this equation can be replaced by

$$C(q^{-1})B(q^{-1}) - D(q^{-1})A(q^{-1}) \equiv 0$$

Using new variables given by

$$\hat{A}(q^{-1}) = 1 + \sum_1^{\hat{n}_a} \hat{a}_i q^{-i} = A(q^{-1}) + C(q^{-1}); \; \hat{n}_a = \max(m_a, n_a)$$

$$\hat{B}(q^{-1}) = \sum_1^{\hat{n}_b} \hat{b}_i q^{-i} = B(q^{-1}) + D(q^{-1}); \; \hat{n}_b = \max(m_b, n_b)$$

the equation is written as

$$\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1}) \equiv 0$$

Let $x^T = [c_1 \ldots c_{m_a} \; d_1 \ldots d_{m_b}]$ and introduce the corresponding operators

$$C(q^{-1}) = \sum_1^{m_a} c_i q^{-i}, \quad D(q^{-1}) = \sum_1^{m_b} d_i q^{-i}$$

Then

$$x^T R x = E\Big[[-y(t-1) \ldots -y(t-m_a) u(t-1) \ldots u(t-m_b)]x\Big]^2$$

$$= E\big[-C(q^{-1})y(t) + D(q^{-1})u(t)\big]^2$$

The equation (2.8) is thus equivalent to

$$\frac{C(q^{-1})B(q^{-1}) - D(q^{-1})A(q^{-1})}{A(q^{-1})} u(t) = 0 \text{ a.s.}$$

From Lemma 2.4 Corr and Lemma 2.6 it follows that this equation can be replaced by

$$C(q^{-1})B(q^{-1}) - D(q^{-1})A(q^{-1}) \equiv 0$$

Using new variables given by

$$\hat{A}(q^{-1}) = 1 + \sum_1^{\hat{n}_a} \hat{a}_i q^{-i} = A(q^{-1}) + C(q^{-1}); \; \hat{n}_a = \max(m_a, n_a)$$

$$\hat{B}(q^{-1}) = \sum_1^{\hat{n}_b} \hat{b}_i q^{-i} = B(q^{-1}) + D(q^{-1}); \; \hat{n}_b = \max(m_b, n_b)$$

the equation is written as

$$\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1}) \equiv 0$$

From Lemma 2.1 itthus follows that:

i) if $n_\ell \leqslant 0$ $x = 0$ is the only solution,

ii) if $n_\ell > 0$ the general solution is given by

$$C(q^{-1}) \equiv A(q^{-1})L(q^{-1})$$

$$D(q^{-1}) \equiv B(q^{-1})L(q^{-1})$$

$$L(q^{-1}) = \sum_1^{n_\ell} \ell_i q^{-i}$$

where the numbers ℓ_i are arbitrary.
As a consequence N(R) has dimension n_ℓ.

Q.E.D.

The following two lemmas were originally used in the author's previous work, Söderström (1972), where also proofs can be found.

Lemma 2.8. Consider the equation

$$F(x) \equiv f(x) + \varepsilon g(x) = 0 \tag{2.9}$$

where dim f = dim g = dim x. Let Ω denote a set with the following properties:

f and g are twice differentiable,
$f(x) = 0$ implies $x = x_0$,
$f'(x_0)$ is non singular.

Then there is a $\varepsilon_1 > 0$ such that $0 < \varepsilon \leqslant \varepsilon_1$ implies that (2.9) has a unique solution $\bar{x}$ in Ω. $\bar{x}$ fulfils

$$\bar{x} - x_0 = 0(\varepsilon), \qquad \varepsilon \to 0$$

Lemma 2.9. Consider the function

$$V(x,y,\varepsilon) = \frac{1}{2} x^T P(y)x + \varepsilon h(x,y)$$

where (x,y) belongs to a set Ω, for which P(y) is a positive definite matrix for all y, twice differentiable with respect to y and h(x,y) a twice differentiable function. ε is considered as a fix parameter.

The following necessary and sufficient conditions for local minimum points in Ω are true.

There is a constant $\varepsilon_0 > 0$ such that if $0 < \varepsilon \leqslant \varepsilon_0$ the following is true.

i) Every stationary point of $V(x,y,\varepsilon)$ in Ω fulfils

$$(x,y) = (0,y_0) + \left[0(\varepsilon),o(1)\right], \quad \varepsilon \to 0 \qquad (2.10)$$

where y_0 is a solution of

$$h'_y(0,y) = 0 \qquad (2.11)$$

If (x,y) is a local minimum point it is necessary that $h''_{yy}(0,y_0)$ is positive definite or positive semidefinite.

ii) If y_0 is a solution of (2.11) and $h''_{yy}(0,y_0)$ is positive definite then there exists a unique local minimum of the form (2.10), and the point will in fact satisfy

$$(x,y) = (0,y_0) + \left[0(\varepsilon),0(\varepsilon)\right], \quad \varepsilon \to 0$$

The matrix of second order derivatives is positive definite in the minimum point.

III. GLOBAL MINIMUM POINTS FOR DIFFERENT STRUCTURES.

In this chapter the global minimum points of loss functions of the type

$$\begin{cases} V(\hat{\theta},\theta) = \frac{1}{2N} \sum\limits_{t=1}^{N} \varepsilon^2(t) \\ \\ \varepsilon(t) = \frac{G(\theta;q^{-1}) - G(\hat{\theta};q^{-1})}{H(\hat{\theta};q^{-1})} u(t) + \frac{H(\theta;q^{-1})}{H(\hat{\theta};q^{-1})} e(t) \end{cases} \tag{3.1}$$

are analyzed.

For finite N the analysis has to be done in a probabilistic setting. In order to do the analysis reasonable ergodic theory will be used.

The following assumptions are made:

o Let Ω = {θ; such that the poles of $G(\theta;z)$, the poles of $H(\theta;z)$ and the zeros of $H(\theta;z)$ are outside the circle $|z| = 1 + \varepsilon$, where $\varepsilon > 0$ is some small number}. It is assumed that $\theta \in \Omega$ and only points $\hat{\theta}$ in the set Ω are considered. This limitation is motivated from the representation theorem, Åström (1970), and the demand of a finite variance of the output.

o The input is assumed to be a periodic signal or filtered white noise (or a sum of these two types).

o The input signal and the noise e(t) are independent.

Under these assumptions it follows from Lemma 2.3 that $V(\hat{\theta},\theta)$ has a limit $W(\hat{\theta},\theta)$ (with probability one and in mean square) as N tends to infinity. The function $W(\hat{\theta},\theta)$ is given by

$$W(\hat{\theta},\theta) = \frac{1}{2}\,\mathcal{E}\left[\frac{G(\theta;q^{-1}) - G(\hat{\theta};q^{-1})}{H(\hat{\theta};q^{-1})}\,u(t)\right]^2 +$$

$$+ \frac{1}{2}\,E\left[\frac{H(\theta;q^{-1})}{H(\hat{\theta};q^{-1})}\,e(t)\right]^2 \tag{3.2}$$

Let

$$\tilde{H}(q^{-1}) = \frac{H(\theta;q^{-1})}{H(\hat{\theta};q^{-1})} = 1 + \sum_{i=1}^{\infty} \tilde{h}_i q^{-i}$$

Then

$$W(\hat{\theta},\theta) \geqslant \frac{1}{2}\,E\left[e(t) + \sum_{i=1}^{\infty} \tilde{h}_i e(t-i)\right]^2 =$$

$$= \frac{1}{2}\,\lambda^2\left[1 + \Sigma\tilde{h}_i^2\right] \geqslant \frac{\lambda^2}{2} \tag{3.3}$$

But $W(\theta,\theta) = \frac{1}{2}\lambda^2$ which implies that $\hat{\theta} = \theta$ always is a global minimum point of $W(\hat{\theta},\theta)$. However, $\hat{\theta} = \theta$ is not necessarily a unique solution of

$$W(\hat{\theta},\theta) = \inf_{\theta^*} W(\theta^*,\theta) \tag{3.4}$$

This equation can in view of (3.3) be written as

$$\begin{cases} \dfrac{G(\theta;q^{-1}) - G(\hat{\theta};q^{-1})}{H(\hat{\theta};q^{-1})}\,u(t) = 0 \\[2ex] H(\theta;q^{-1}) = H(\hat{\theta};q^{-1}) \end{cases} \tag{3.5}$$

The equations (3.5) will now be discussed for different structures of the system. The input signal will be assumed to be persistently exciting of a sufficiently high order. The first part of (3.5) will then in fact be replaced by

$$G(\hat{\theta};q^{-1}) = G(\theta;q^{-1})$$

Most of the material is well-known and parts of it have been treated by the author before in Söderström (1972), Åström-Söderström (1973). These parts are included here to get a more complete survey.

As a general result it can be said that the loss functions for the different cases have a unique global minimum if a model of correct order is applied. If the model order is too high there is in most cases no unique global minimum point.

To simplify the notations the second argument in W will be dropped in the rest of the report.

Structure 1: The Least Squares (LS) Method.

The system is in this case given by

$$A(q^{-1})y(t) = B(q^{-1})u(t) + e(t) \tag{3.6}$$

so

$$G(\theta;q^{-1}) = \frac{B(q^{-1})}{A(q^{-1})} \qquad H(\theta;q^{-1}) = \frac{1}{A(q^{-1})}$$

The equations (3.5) become

$$\begin{cases} \dfrac{\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1})}{A(q^{-1})}\, u(t) = 0 \\[2ex] \hat{A}(q^{-1}) \equiv A(q^{-1}) \end{cases}$$

or simplified

$$[\hat{B}(q^{-1}) - B(q^{-1})]u(t) = 0$$

$$\hat{A}(q^{-1}) \equiv A(q^{-1}) \tag{3.7}$$

The consistency properties of this method are well-known, Åström (1968).

Lemma 3.1. Assume that $n_\ell = \min(\hat{n}_a - n_a, \hat{n}_b - n_b) \geq 0$ and that u(t) is persistently exciting of order $\hat{n}_b$. Then there is a unique global minimum point given by $\hat{A}(q^{-1}) \equiv (A(q^{-1})$, $\hat{B}(q^{-1}) \equiv B(q^{-1})$. There are no other local minimum points.

Proof. The first statement follows immediately from Lemma 2.5 and (3.7). The second statement is true since V is convex.

Q.E.D.

Structure 2: The General Least Squares (GLS) Model.

The structure is given by Clarke (1967), Söderström (1972)

$$A(q^{-1})y(t) = B(q^{-1})u(t) + \frac{1}{C(q^{-1})}\, e(t) \tag{3.8}$$

Thus

$$G(\theta;q^{-1}) = \frac{B(q^{-1})}{A(q^{-1})}, \qquad H(\theta;q^{-1}) = \frac{1}{A(q^{-1})C(q^{-1})}$$

The equations (3.5) become

$$\begin{cases} \dfrac{\hat{C}(q^{-1})}{A(q^{-1})} [\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1})]u(t) = 0 \\ \\ \hat{A}(q^{-1})\hat{C}(q^{-1}) \equiv A(q^{-1})C(q^{-1}) \end{cases} \tag{3.9}$$

The solution of these equations is treated in Söderström (1972).

Lemma 3.2. Assume that $n_\ell = \min(\hat{n}_a - n_a, \hat{n}_b - n_b) \geq 0$, $(\hat{n}_c - n_c) \geq 0$, u(t) is persistently exciting of order $\max(\hat{n}_a + n_b, n_a + \hat{n}_b)$, and that $A(q^{-1})$ and $B(q^{-1})$ are relatively prime. Then the solutions of (3.9) fulfil

$$\begin{aligned} \hat{A}(q^{-1}) &\equiv A(q^{-1})L(q^{-1}) \\ \hat{B}(q^{-1}) &\equiv B(q^{-1})L(q^{-1}) \\ L(q^{-1})\hat{C}(q^{-1}) &\equiv C(q^{-1}) \end{aligned} \tag{3.10}$$

where

$$\begin{aligned} L(q^{-1}) &= 1 + \sum_{1}^{n_\ell} \ell_i q^{-i} \quad \text{if } n_\ell \geq 1 \\ \\ &= 1 \qquad\qquad\quad\ \ \text{if } n_\ell = 0 \end{aligned}$$

Proof. The assumptions of the theorem imply that the first equation in (3.9) can be replaced by

$$\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1}) \equiv 0$$

Lemma 2.1 gives the rest of the proof.

Q.E.D.

Remark 1. If $n_\ell = 0$ $\hat{\theta} = \theta$ is the unique solution.

Remark 2. Note that when $n_\ell \geqslant 1$ there are only a finite number of solutions of (3.9). This is particular for the GLS case. The reason for this property is the special structure of the system equation.

Remark 3. If u(t) is not persistently exciting of order $\max(\hat{n}_a+n_b,\ n_a+\hat{n}_b)$ there may exist global minimum points which do not fulfil (3.10). An example is given in Söderström (1972).

It is well-known, Söderström (1972), that the number of local minimum points depends on the signal to noise ratio.

Structure 3: Time Series.

In this case stochastic processes of the form

$$A(q^{-1})y(t) = C(q^{-1})e(t) \tag{3.11}$$

are considered. Then

$$G(\theta;q^{-1}) = 0 \qquad\qquad H(\theta;q^{-1}) = \frac{C(q^{-1})}{A(q^{-1})}$$

The equations for the global minimum point (3.5) are

$$A(q^{-1})\hat{C}(q^{-1}) - \hat{A}(q^{-1})C(q^{-1}) \equiv 0 \tag{3.12}$$

Lemma 3.3. Assume that

i) $A(q^{-1})$ and $C(q^{-1})$ are relatively prime

ii) $n_\ell = \min(\hat{n}_a - n_a, \hat{n}_c - n_c) \geqslant 0$

The solutions of (3.12) are

$$\begin{aligned} \hat{A}(q^{-1}) &\equiv A(q^{-1})L(q^{-1}) \\ \hat{C}(q^{-1}) &= C(q^{-1})L(q^{-1}) \end{aligned} \tag{3.13}$$

where

$$\begin{aligned} L(q^{-1}) &= 1 + \sum_{1}^{n_\ell} \ell_i q^{-i} \quad \text{if } n_\ell \geqslant 1 \\ &= 1 \qquad\qquad\qquad \text{if } n_\ell = 0 \end{aligned}$$

The parameters ℓ_i are arbitrary. These points are the only stationary points as well.

Proof. See Åström-Söderström (1973).

Remark. If $n_\ell = 0$ $\hat{\theta} = \theta$ is the only global as well as local minimum point.

Structure 4.

The system is assumed to be governed by

$$A(q^{-1})y(t) = B(q^{-1})u(t) + A(q^{-1})e(t) \tag{3.14}$$

so

$$G(\theta;q^{-1}) = \frac{B(q^{-1})}{A(q^{-1})} \qquad H(\theta;q^{-1}) = 1$$

The equations (3.5) are thus replaced by

$$\frac{[\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1})]}{A(q^{-1})\hat{A}(q^{-1})} u(t) = 0 \text{ a.s.} \tag{3.15}$$

Lemma 3.4. Assume that $n_\ell = \min(\hat{n}_a - n_a, \hat{n}_b - n_b) \geq 0$, $A(q^{-1})$ and $B(q^{-1})$ are relatively prime, and that $u(t)$ is persistently exciting of order $\max(\hat{n}_a + n_b, n_a + \hat{n}_b)$. Then the solutions of (3.15) are

$$\begin{aligned} \hat{A}(q^{-1}) &\equiv A(q^{-1})L(q^{-1}) \\ \hat{B}(q^{-1}) &\equiv B(q^{-1})L(q^{-1}) \end{aligned} \tag{3.16}$$

where

$$L(q^{-1}) = 1 + \sum_{1}^{n_\ell} \ell_i q^{-i} \qquad n_\ell \geq 1$$

$$= 1 \qquad n_\ell = 0$$

The numbers ℓ_i are arbitrary.

Proof. Lemma 2.4 and Lemma 2.1 give the result.

Q.E.D.

Remark. If $n_\ell = 0$ $\hat{\theta} = \theta$ is the only global minimum point.

Structure 5.

This structure is discussed e.g. in Åström-Bohlin (1966). It is given by

$$A(q^{-1})y(t) = B(q^{-1})u(t) + C(q^{-1})e(t) \tag{3.17}$$

This means that

$$G(\theta;q^{-1}) = \frac{B(q^{-1})}{A(q^{-1})} \qquad H(\theta;q^{-1}) = \frac{C(q^{-1})}{A(q^{-1})}$$

and (3.5) can be replaced by

$$\begin{cases} \dfrac{\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1})}{A(q^{-1})\hat{C}(q^{-1})} u(t) = 0 \\ \\ \hat{A}(q^{-1})C(q^{-1}) - A(q^{-1})\hat{C}(q^{-1}) \equiv 0 \end{cases} \tag{3.18}$$

Lemma 3.5. Assume that $n_\ell = \min(\hat{n}_a - n_a,\ \hat{n}_b - n_b,\ \hat{n}_c - n_c) \geq 0$, $u(t)$ is persistently exciting of order $\max(\hat{n}_a + n_b,\ n_a + \hat{n}_b)$, and that $A(q^{-1})$, $B(q^{-1})$ and $C(q^{-1})$ are relatively prime. Then the general solution of (3.18) is given by

$$\begin{aligned} \hat{A}(q^{-1}) &\equiv A(q^{-1})L(q^{-1}) \\ \hat{B}(q^{-1}) &\equiv B(q^{-1})L(q^{-1}) \\ \hat{C}(q^{-1}) &\equiv C(q^{-1})L(q^{-1}) \end{aligned} \tag{3.19}$$

where

$$\begin{aligned} L(q^{-1}) &= 1 + \sum_{1}^{n_\ell} \ell_i q^{-i} && \text{if } n_\ell \geq 1 \\ &= 1 && \text{if } n_\ell = 0 \end{aligned}$$

The coefficients ℓ_i are arbitrary.

Proof. Define $\tilde{A}(q^{-1})$, $\tilde{B}(q^{-1})$ and $D(q^{-1})$ from

$$A(q^{-1}) = \tilde{A}(q^{-1})D(q^{-1})$$

$$B(q^{-1}) = \tilde{B}(q^{-1})D(q^{-1})$$

$\tilde{A}(q^{-1})$, $\tilde{B}(q^{-1})$ are relatively prime

$$D(q^{-1}) = 1 + \sum_{1}^{n_d} d_i q^{-i} \qquad (n_d \geqslant 0)$$

The first equation of (3.18) can be replaced by (Lemma 2.4)

$$\hat{A}(q^{-1})\tilde{B}(q^{-1}) - \tilde{A}(q^{-1})\hat{B}(q^{-1}) \equiv 0$$

The solution is (Lemma 2.1)

$$\hat{A}(q^{-1}) \equiv \tilde{A}(q^{-1})M(q^{-1})$$

$$\hat{B}(q^{-1}) \equiv \tilde{B}(q^{-1})M(q^{-1}) \qquad (3.20)$$

$$M(q^{-1}) = 1 + \sum_{1}^{n_m} m_i q^{-i}$$

$$n_m = n_d + n_\ell$$

$(m_i)_{i=1}^{n_m}$ are determined only by the second equation in (3.18)

The last equation of (3.18) gives

$$M(q^{-1})C(q^{-1}) - D(q^{-1})\hat{C}(q^{-1}) \equiv 0 \qquad (3.21)$$

According to the assumptions of the lemma $C(q^{-1})$ and $D(q^{-1})$ have no common factors.

The solution of (3.21) w.r.t. $M(q^{-1})$ and $\hat{C}(q^{-1})$ is

$$\begin{aligned} \hat{C}(q^{-1}) &\equiv C(q^{-1})L(q^{-1}) \\ M(q^{-1}) &\equiv D(q^{-1})L(q^{-1}) \\ L(q^{-1}) &= 1 + \sum_{1}^{n_\ell} \ell_i q^{-i} \end{aligned} \tag{3.22}$$

$(\ell_i)_1^{n_\ell}$ arbitrary

The combination of (3.20) and (3.22) gives the desired solution (3.19).

Q.E.D.

Remark. If $n_\ell = 0$ $\hat{\theta} = \theta$ is the only global minimum point.

Structure 6.

In this section the structure used by Bohlin (1970) is considered

$$y(t) = \frac{B(q^{-1})}{A(q^{-1})} u(t) + \frac{C(q^{-1})}{D(q^{-1})} e(t) \tag{3.23}$$

The equations (3.5) turn out to be

$$\begin{cases} \dfrac{\hat{D}(q^{-1})}{\hat{C}(q^{-1})} \dfrac{\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1})}{A(q^{-1})\hat{A}(q^{-1})} u(t) = 0 \text{ a.s.} \\ \hat{C}(q^{-1})D(q^{-1}) \equiv C(q^{-1})\hat{D}(q^{-1}) \end{cases} \tag{3.24}$$

Lemma 3.6. Assume that

$n_\ell = \min(\hat{n}_a - n_a,\ \hat{n}_b - n_b) \geq 0$

$n_m = \min(\hat{n}_c - n_c,\ \hat{n}_d - n_d) \geq 0$

$A(q^{-1})$ and $B(q^{-1})$ are relatively prime

$C(q^{-1})$ and $D(q^{-1})$ are relatively prime

$u(t)$ persistently exciting of order $\max(\hat{n}_a + n_b,\ n_a + \hat{n}_b)$

Then the general solution of (3.24) is

$$\begin{aligned}
\hat{A}(q^{-1}) &\equiv A(q^{-1})L(q^{-1})\\
\hat{B}(q^{-1}) &\equiv B(q^{-1})L(q^{-1})\\
\hat{C}(q^{-1}) &\equiv C(q^{-1})M(q^{-1})\\
\hat{D}(q^{-1}) &\equiv D(q^{-1})M(q^{-1})
\end{aligned} \qquad (3.25)$$

where

$$L(q^{-1}) = 1 + \sum_1^{n_\ell} \ell_i q^{-i}$$

$$M(q^{-1}) = 1 + \sum_1^{n_m} m_i q^{-i}$$

$(\ell_i)_1^{n_\ell}$, $(m_i)_1^{n_m}$ arbitrary

Proof. The result follows from Lemma 2.1 and Lemma 2.4.

Q.E.D.

Remark. If $n_\ell = n_m = 0$, $\hat{\theta} = \theta$ is the only global minimum point.

IV. LOCAL MINIMUM POINTS FOR STRUCTURE 4.

In this chapter the local minimum points for the case with white measurements noise are treated. It will be shown that $n_a = \hat{n}_a = 1$, n_b arbitrary will imply a unique local minimum point. The loss function can in certain cases have "singular" saddle points corresponding to $\hat{B}(q^{-1}) \equiv 0$. The analysis can unfortunately not be extended to the case $n_a > 1$.

For the structure with white measurement noise the system is described by

$$A(q^{-1})y(t) = B(q^{-1})u(t) + A(q^{-1})e(t)$$

The loss function for this structure is given by

$$2W(\hat{\theta}) = \mathcal{E}\left[\frac{\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1})}{A(q^{-1})\hat{A}(q^{-1})}\, u(t)\right]^2 + \lambda^2 \qquad (4.1)$$

Assume that $\hat{n}_a \geq n_a$, $\hat{n}_b \geq n_b$, $B(q^{-1}) \not\equiv 0$ and that $u(t)$ is persistently exciting of order $\max(\hat{n}_a + n_b,\ n_a + \hat{n}_b)$.

The stationary points of the function are the solutions of $W_{\hat{\theta}}(\hat{\theta}) = 0$, which is written as

$$\mathcal{E}\left[\frac{\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1})}{A(q^{-1})\hat{A}(q^{-1})}\, u(t)\right]\left[\frac{\hat{B}(q^{-1})}{\hat{A}(q^{-1})^2}\, q^{-i}u(t)\right] = 0 \qquad 1 \leq i \leq \hat{n}_a$$

$$(4.2)$$

$$\mathcal{E}\left[\frac{\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1})}{A(q^{-1})\hat{A}(q^{-1})}\, u(t)\right]\left[\frac{-1}{\hat{A}(q^{-1})}\, q^{-i}u(t)\right] = 0 \qquad 1 \leq i \leq \hat{n}_b$$

It is not possible to find all solutions of (4.2) in an easy way. The following attempt of analysis will be made. Let

$$H(q^{-1}) = \sum_{1}^{m} h_i q^{-i} = \hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1})$$

with $m = \max(\hat{n}_a+n_b,\ n_a+\hat{n}_b)$. The equations (4.2) will be rewritten as

$$Q(\hat{\theta})\begin{bmatrix} h_1 \\ \vdots \\ h_m \end{bmatrix} = 0$$

where $Q(\hat{\theta})$ is a matrix of order $(\hat{n}_a+\hat{n}_b) \times m$. If rank $Q(\hat{\theta})$ is m for all $\hat{\theta}$ it can be concluded that $h_i = 0$ for i = 1, ..., m. This gives the equation for the global minimum points, Lemma 3.4.

Put

$$v(t) = \frac{1}{A(q^{-1})\hat{A}(q^{-1})^2}\, u(t)$$

Then (4.2) is equivalent to

$$\mathcal{E}\begin{bmatrix} -\hat{B}(q^{-1})A(q^{-1})v(t-1) \\ \vdots \\ -\hat{B}(q^{-1})A(q^{-1})v(t-\hat{n}_a) \\ \hat{A}(q^{-1})A(q^{-1})v(t-1) \\ \vdots \\ \hat{A}(q^{-1})A(q^{-1})v(t-\hat{n}_b) \end{bmatrix}\begin{bmatrix} \hat{A}(q^{-1})v(t-1) & \ldots & \hat{A}(q^{-1})v(t-m) \end{bmatrix}\begin{bmatrix} h_1 \\ \cdot \\ \cdot \\ \cdot \\ h_m \end{bmatrix} = 0$$

which can be written as

$$\left[\begin{array}{cccccccc} 0 & -\hat{b}_1 & \cdot & \cdot & -\hat{b}_{\hat{n}_b} & & 0 & \\ & \ddots & & \ddots & & \ddots & & \\ 0 & & & 0 & -\hat{b}_1 & \cdot & \cdot & -\hat{b}_{\hat{n}_b} \\ \hdashline 1 & \cdot & \cdot & \cdot & \hat{a}_{\hat{n}_a} & & 0 & \\ & \ddots & & & & \ddots & & \\ 0 & & & 1 & \cdot & \cdot & \cdot & \hat{a}_{\hat{n}_a} \end{array}\right] \cdot P_0 \cdot \begin{bmatrix} h_1 \\ \cdot \\ \cdot \\ \cdot \\ h_m \end{bmatrix} = 0 \qquad (4.3)$$

where P_0 is the following matrix of order $(\hat{n}_a+\hat{n}_b) \times m$.

$$P_0 = \mathcal{E}\begin{bmatrix} A(q^{-1})v(t-1) \\ \vdots \\ A(q^{-1})v(t-\hat{n}_a-\hat{n}_b) \end{bmatrix}\left[\hat{A}(q^{-1})v(t-1) \ \ldots \ \hat{A}(q^{-1})v(t-m)\right] \qquad (4.4)$$

To continue the analysis it is necessary to examine the rank of the two matrices in (4.3). It will be necessary to separate three different cases.

Case 1.

Consider points such that $\hat{A}(q^{-1})$ and $\hat{B}(q^{-1})$ are relatively prime. Then the first matrix of (4.3) is non singular (Lemma 2.2). Define a square matrix P of order m × m.

$$P(A,\hat{A},v) = \mathcal{E}\begin{bmatrix} A(q^{-1})v(t-1) \\ \vdots \\ A(q^{-1})v(t-m) \end{bmatrix}\left[\hat{A}(q^{-1})v(t-1) \ \ldots \ \hat{A}(q^{-1})v(t-m)\right] \qquad (4.5)$$

which consists of the upper square part of P_0.

If P is non singular for all possible $(\hat{a}_i)_{i=1}^{\hat{n}_a}$ then it is possible to conclude that $h_i = 0$ is the only solution of (4.3).

The properties of $P(A,\hat{A},v)$ are described in

Lemma 4.1.

i) Assume that $n_a = 1$, $\hat{n}_a = 1$. Then $P(A,\hat{A},v)$ is non singular for all A, all $\hat{A}$ and all v(t), such as v(t) is persistently exciting of order m.

ii) There are A, $\hat{A}$ and v(t) such that $n_a = 1$, $\hat{n}_a = 2$, $m = 3$, v(t) persistently exciting of order m and $P(A,\hat{A},v)$ singular.

Proof.

i) Let $x = [x_1 \; \dots \; x_m]^T$ be an arbitrary vector and define

$$X(q^{-1}) = \sum_1^m x_i q^{-i}$$

Then

$$x^T Px = \mathcal{E}\,[A(q^{-1})X(q^{-1})v(t)][\hat{A}(q^{-1})X(q^{-1})v(t)] =$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{Re}[A(e^{i\omega})\hat{A}(e^{-i\omega})]\,|X(e^{i\omega})|^2 \phi_v(\omega)d\omega$$

The function $\phi_v(\omega)$ is the spectral density associated with the asymptotic sample covariance function.

But

$$\mathrm{Re}[A(e^{i\omega})\hat{A}(e^{-i\omega})] =$$

$$= 1 + a\hat{a} + (a+\hat{a})\cos\omega \geqslant 1 + a\hat{a} - |a+\hat{a}| \geqslant$$

$$\geqslant (1-|a|)(1-|\hat{a}|) > 0$$

Thus $x^T Px \geqslant 0$ and equality implies $|X(e^{i\omega})|^2\phi_v(\omega) \equiv \equiv 0$. From Lemma 2.4 it is concluded that this implies $X(e^{i\omega}) \equiv 0$ or $x = 0$.

ii) Let $v(t)$ be white noise with unit variance. Take $A(q^{-1}) = 1 + aq^{-1}$ and $\hat{A}(q^{-1}) = (1-aq^{-1})^2$. Then

$$P(A,\hat{A},v) = \begin{bmatrix} 1-2a^2 & a & 0 \\ -2a+a^3 & 1-2a^2 & a \\ a^2 & -2a+a^3 & 1-2a^2 \end{bmatrix}$$

$\det P = 1 - 2a^2 + 3a^4 - 4a^6$ is a continuous function of a. $\mathrm{Det}[P(a=0)] = 1$ and $\det[P(a=1)] = -2$ imply that there is $|a| < 1$ such that $\det P = 0$.

Q.E.D.

Case 2.

Consider points such that $\hat{A}(q^{-1})$ and $\hat{B}(q^{-1})$ are not relatively prime, but $\hat{B}(q^{-1}) \not\equiv 0$.

Define

$$\bar{A}(q^{-1}) = 1 + \sum_1^{\bar{n}_a} \bar{a}_i q^{-i}; \quad \bar{B} = \sum_1^{\bar{n}_b} \bar{b}_i q^{-i}; \quad \bar{L}(q^{-1}) = 1 + \sum_1^{\bar{n}_\ell} \bar{\ell}_i q^{-i}$$

by

$$\hat{A}(q^{-1}) = \bar{A}(q^{-1})\bar{L}(q^{-1})$$

$$\hat{B}(q^{-1}) = \bar{B}(q^{-1})\bar{L}(q^{-1})$$

$\bar{A}(q^{-1})$ and $\bar{B}(q^{-1})$ are relatively prime

$$\bar{n}_a = \hat{n}_a - \bar{n}_\ell ; \; \bar{n}_b = \hat{n}_b - \bar{n}_\ell$$

Change the definition of $H(q^{-1})$, m and v(t) to

$$H(q^{-1}) = \sum_1^m h_i q^{-i} = \bar{A}(q^{-1})B(q^{-1}) - A(q^{-1})\bar{B}(q^{-1})$$

$$m = \max(\bar{n}_a+n_b , \; n_a+\bar{n}_b)$$

$$v(t) = \frac{1}{A(q^{-1})\bar{A}(q^{-1})\hat{A}(q^{-1})} u(t)$$

Then the equation $V_{\hat{\theta}}(\hat{\theta}) = 0$ can be written as

$$\begin{array}{r} n_{\hat{a}} \left\{ \begin{array}{c} \\ \\ \\ \end{array} \right. \\ n_{\hat{b}} \left\{ \begin{array}{c} \\ \\ \\ \end{array} \right. \end{array}
\left[\begin{array}{cccccccc}
0 & -\bar{b}_1 & \cdot & \cdot & -\bar{b}_{\bar{n}_b} & & & 0 \\
 & \ddots & & & & \ddots & & \\
 & & \ddots & & & & \ddots & \\
0 & & & 0 & -\bar{b}_1 & \cdot & \cdot & -\bar{b}_{\bar{n}_b} \\
\hline
1 & \bar{a}_1 & \cdot & \cdot & \bar{a}_{\bar{n}_a} & & 0 & \\
 & \ddots & & & & \ddots & & \\
 & 0 & \ddots & & & & \ddots & \\
 & & & 1 & \bar{a}_1 & \cdot & \cdot & \bar{a}_{\bar{n}_a}
\end{array} \right] \cdot Q_0 \cdot \begin{bmatrix} h_1 \\ \cdot \\ \cdot \\ \cdot \\ h_m \end{bmatrix} = 0 \quad (4.6)$$

$\hat{n}_a+\hat{n}_b-\bar{n}_\ell$ columns

where Q_0 is the following matrix of order $(\hat{n}_a+\hat{n}_b-\bar{n}_\ell) \times m$

$$Q_0 = \mathcal{E}\begin{bmatrix} A(q^{-1})v(t-1) \\ \vdots \\ A(q^{-1})v(t-\hat{n}_a-\hat{n}_b+\bar{n}_\ell) \end{bmatrix}\begin{bmatrix}\hat{A}(q^{-1})v(t-1) & \ldots & \hat{A}(q^{-1})v(t-m)\end{bmatrix} \tag{4.7}$$

According to Lemma 2.2 the first matrix of (4.6) has rank $\hat{n}_a + \hat{n}_b - \bar{n}_\ell$. Thus $h_i = 0$ is the only solution if rank $Q_0 =$ $= m$. This condition, however, is already analyzed in the previous case.

Case 3.

Consider points such that $\hat{B}(q^{-1}) \equiv 0$. Such singular points may look uninteresting from a theoretical point of view. For two reasons they are studied here, besides the purpose to give general information of the loss function W. The first reason is that in a practical case it is not trivial to determine if $\hat{b}_i = 0$. The other reason is that the result of this chapter will be used later on in Chapter 6. For this case the equations (4.2) turn out to be

$$\mathcal{E}\left[\frac{B(q^{-1})}{A(q^{-1})}\ u(t)\right]\left[\frac{q^{-i}}{\hat{A}(q^{-1})}\ u(t)\right] = 0 \qquad 1 \leqslant i \leqslant \hat{n}_b \tag{4.8}$$

If $\hat{n}_a > \hat{n}_b$ this system is overdetermined and may have an infinite number of solutions such that $\hat{A}(z)$ has zeros outside the unit circle. In Appendix A this case is further considered. It is also shown that the stationary points satisfying $\hat{B}(q^{-1}) \equiv 0$ always are saddle points.

The equation (4.8) implies

$$\mathcal{E}\left[\frac{B(q^{-1})}{A(q^{-1})}\ u(t-1)\right]\left[\frac{B(q^{-1})}{\hat{A}(q^{-1})}\ u(t-1)\right] = 0 \tag{4.9}$$

Put

$$v(t) = \frac{B(q^{-1})}{A(q^{-1})\hat{A}(q^{-1})} u(t)$$

If $n_a = \hat{n}_a = 1$ it follows from Lemma 4.1 that (4.9) cannot be satisfied.

In Appendix B the special case of u(t) as white noise is treated. It is shown that the mild condition $\hat{n}_a = \hat{n}_b \geq$ $\geq \max(n_a, n_b)$ implies that the global minimum points are the only stationary points.

Summing up, the analysis has given the following information of the loss function $W(\hat{\theta})$. Assume that u(t) is persistently exciting of order $\max(\hat{n}_a + n_b,\ n_a + \hat{n}_b)$

1. If $\hat{n}_a = n_a = 1$, $\hat{n}_b \geq n_b$, n_b arbitrary then $W(\hat{\theta})$ has a unique stationary point, namely the local (and global) minimum point $\hat{\theta} = \theta$.

2. If $\hat{n}_a > n_a$ and $\hat{n}_b > n_b$ there is no unique global minimum point.

3. The analysis gives no information of the number of local minimum points if $\hat{n}_a \geq 2$.

4. There are systems such that $\hat{n}_a = n_a = 2$, $\hat{n}_b = n_b = 1$ and with a set of saddle points satisfying $\hat{B}(q^{-1}) \equiv 0$. An immediate implication is that it is not sufficient to consider only $W'(\hat{\theta}) = 0$ in the general analysis of the number of local minimum points.

5. If u(t) is white noise and $\hat{n}_a = \hat{n}_b \geq \max(n_a, n_b)$ then the global minimum points are the only stationary points.

V. LOCAL MINIMUM POINT FOR STRUCTURE 5.

In this part structure 5 is considered. Partial results on the number of local minimum points will be given. Only cases with very high or very low signal to noise ratios are treated. The mathematical tools are Lemma 2.8 and Lemma 2.9. These two lemmas deal with the effects of a disturbance term $\varepsilon g(x)$ resp. $\varepsilon h(x,y)$. The application of them will be made on the loss function. ε will be inverse proportional or proportional to the signal to noise ratio.

Theorem 5.1. Consider the system

$$A(q^{-1})y(t) = B(q^{-1})u(t) + C(q^{-1})e(t) \qquad (3.17) = (5.1)$$

and the loss function

$$2W(\hat{\theta}) = E\left[\frac{\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1})}{A(q^{-1})\hat{C}(q^{-1})}\,u(t)\right]^2 +$$

$$+ E\left[\frac{\hat{A}(q^{-1})C(q^{-1})}{A(q^{-1})\hat{C}(q^{-1})}\,e(t)\right]^2 \qquad (5.2)$$

Assume that

i) $n_\ell = \min(\hat{n}_a - n_a,\ \hat{n}_b - n_b,\ \hat{n}_c - n_c) = 0$

ii) $u(t)$ is persistently exciting of order $\max(\hat{n}_a + n_b,\ n_a + \hat{n}_b)$

iii) $A(q^{-1})$, $B(q^{-1})$ and $C(q^{-1})$ are relatively prime.

Denote the signal to noise ratio by S. There is a number S_0 (which may depend on Ω), such that if $S_0 \leq S < \infty$ then

$W(\hat{\theta})$ has a unique local minimum in Ω, namely $\hat{\theta} = \theta$.

Proof. This proof is a modification of Appendix E in Söderström (1972). Perform a change of variables by

$$x = \begin{bmatrix} \hat{a}_1 - a_1 \\ \vdots \\ \hat{a}_{n_a} - a_{n_a} \\ \vdots \\ \hat{a}_{\hat{n}_a} \\ \hat{b}_1 - b_1 \\ \vdots \\ \hat{b}_{n_b} - b_{n_b} \\ \vdots \\ \hat{b}_{\hat{n}_b} \end{bmatrix} \qquad y = \begin{bmatrix} \hat{c}_1 \\ \cdot \\ \cdot \\ \cdot \\ \hat{c}_{\hat{n}_c} \end{bmatrix} \tag{5.3}$$

Assume that $A(q^{-1}) \equiv \tilde{A}(q^{-1})D(q^{-1})$, $B(q^{-1}) \equiv \tilde{B}(q^{-1})D(q^{-1})$ where $\tilde{A}(q^{-1})$ and $\tilde{B}(q^{-1})$ are relatively prime and

$$D(q^{-1}) = 1 + \sum_1^{n_d} d_i q^{-i} \qquad (n_d \geq 0)$$

The loss function can be written

$$W(x,y) = \frac{1}{2}\, x^T P(y) x + \varepsilon \tilde{h}(x,y) \tag{5.4}$$

with $P(y)$ as the system covariance matrix of

$$A(q^{-1})y^F(t) = B(q^{-1})u^F(t), \qquad u^F(t) = \frac{1}{\hat{C}(q^{-1})}\, u(t)$$

where $u^F(t)$ is the input and $y^F(t)$ the output.

P(y) may be singular, but the null space of P(y) is independent of y. This is obvious, since from Lemma 2.7 the null space is spanned by vectors of the form

$$\begin{bmatrix} f_1 \\ \vdots \\ f_{\hat{n}_a} \\ g_1 \\ \vdots \\ g_{\hat{n}_b} \end{bmatrix}$$

with

$$F(q^{-1}) = \sum_1^{\hat{n}_a} f_i q^{-i} \equiv \tilde{A}(q^{-1})L'(q^{-1})$$

$$G(q^{-1}) = \sum_1^{\hat{n}_b} g_i q^{-i} = \tilde{B}(q^{-1})L'(q^{-1})$$

$$L'(q^{-1}) = \sum_1^{k} \ell'_i q^{-i} \quad \text{is arbitrary}$$

The simplest case k = 0 is not treated explicitly in the following. In this case P(y) is non singular. It is easy to see how the proof can be simplified for this case.

Introduce now the new variables

$$x' = \begin{bmatrix} x'_1 \\ \\ x'_2 \end{bmatrix}$$

where x_1' is of dimension k and x_2' of dimension $(\hat{n}_a+\hat{n}_b-k)$. The vector x' is defined by

$$x = Qx' = [Q_1 \mid Q_2] \begin{bmatrix} x_1' \\ \hdashline x_2' \end{bmatrix}$$

where

$$Q_1 = \left[\begin{array}{ccc}
1 & & 0 \\
\tilde{a}_1 & & \\
\vdots & \ddots & \\
\tilde{a}_{\tilde{n}_a} & & 1 \\
 & \ddots & \vdots \\
0 & & \tilde{a}_{\tilde{n}_a} \\
 & & 0 \\
\hline
0 & & 0 \\
\tilde{b}_1 & & \\
\vdots & \ddots & \\
\tilde{b}_{\tilde{n}_b} & & \tilde{b}_1 \\
 & \ddots & \vdots \\
0 & & \tilde{b}_{\tilde{n}_b} \\
 & & 0
\end{array}\right]$$

Q_1 is a $(\hat{n}_a+\hat{n}_b) \times k$ matrix and Q_2 an arbitrary $(\hat{n}_a+\hat{n}_b) \times$ $\times$ $(\hat{n}_a+\hat{n}_b-k)$ matrix with the properties $Q_1^TQ_2 = 0$ and Q non singular. Q_2 can for instance be constructed by Gram Schmidt orthogonalization.

From the discussion it follows that

Q_1x_1' is a typical element in the null space $N(P(y))$

Q_2x_2' is a typical element in the space $N(P(y))^{\perp}$

From these facts it is concluded that

$$P(y)Q_1 = 0$$

and that the matrix

$$R(y) = Q_2^TP(y)Q_2$$

of order $(\hat{n}_a+\hat{n}_b-k) \times (\hat{n}_a+\hat{n}_b-k)$ is non singular for all y.

The loss function is now written as

$$W(x_2',z) = \frac{1}{2}\, x_2'^TR(z)x_2' + \varepsilon h(x_2',z) \tag{5.5}$$

where z denotes the vector

$$\begin{bmatrix} x_1' \\ y \end{bmatrix}$$

Write the vector x_1' as

$$x_1' = \begin{bmatrix} \ell_1 \\ \vdots \\ \ell_k \end{bmatrix}$$

Then $x = Q_1x_1'$ is equivalently expressed as

$$\hat{A}(q^{-1}) \equiv A(q^{-1})\hat{L}(q^{-1}), \quad \hat{B}(q^{-1}) \equiv B(q^{-1})\hat{L}(q^{-1})$$

with

$$\hat{L}(q^{-1}) = 1 + \hat{\ell}_1 q^{-1} + \ldots + \hat{\ell}_k q^{-k}$$

The function $h(0,z)$ is written with operators as

$$h(0,z) = E\left[\frac{\hat{L}(q^{-1})C(q^{-1})}{D(q^{-1})\hat{C}(q^{-1})}\, e(t)\right] \tag{5.6}$$

From assumption iii) and the discussion above it follows that $k \geqslant n_d$, $\min(k-n_d, \hat{n}_c-n_c) = 0$ and that $C(q^{-1})$ and $D(q^{-1})$ are relatively prime. From Lemma 3.3 it follows that $h(0,z)$ has a unique local minimum point given by

$$\hat{L}(q^{-1}) \equiv D(q^{-1})$$

$$\hat{C}(q^{-1}) \equiv C(q^{-1})$$

The matrix of second order derivatives of h in this point turns out to be the system covariance matrix of the system

$$y'(t) = \frac{C(q^{-1})}{D(q^{-1})}\, u'(t); \qquad u'(t) = \frac{1}{C(q^{-1})}\, e(t)$$

It is positive definite according to Lemma 2.7.

From Lemma 2.9 it follows that V has a unique local minimum point in Ω. It fulfils

$$\hat{\theta} = \theta + \theta(1/S) \qquad S_0 \leqslant S \to \infty \tag{5.7}$$

Since $\hat{\theta} = \theta$ is a minimum point it is concluded that it is the only local minimum point in Ω.

Q.E.D.

The other theorem of this chapter deals with the case of a low signal to noise ratio and utilizes Lemma 2.8.

Theorem 5.2. Consider the system (5.1) and the loss function (5.2).

Assume that

i) $\min(\hat{n}_a - n_a, \hat{n}_c - n_c) = 0$

ii) u(t) is persistently of order $\hat{n}_b$

iii) $A(q^{-1})$ and $C(q^{-1})$ are relatively prime.

Denote the signal to noise ratio by S.

There is a number S_1 such that $0 < S \leqslant S_1$ implies that $W(\hat{\theta})$ has a unique local minimum in Ω, namely $\hat{\theta} = \theta$.

Proof. From the equation

$$\frac{\partial W}{\partial \hat{b}_i} = 0 \qquad 1 \leqslant i \leqslant \hat{n}_b$$

$\{\hat{b}_i\}$ can be solved as functions of $\{\hat{a}_i, \hat{c}_j\}$ according to assumption ii). Cf. the representation (5.4) of the loss function. These $\hat{b}_i$ functions are put into the remaining equations. The remaining equations can be written (after division by λ^2)

$$f(x) + \varepsilon g(x) = 0 \tag{5.8}$$

x is the vector $[\hat{a}_1 \ldots \hat{a}_{\hat{n}_a}, \hat{c}_1 \ldots \hat{c}_{\hat{n}_c}]^T$. f(x) is the gradient of

$$E\left[\frac{\hat{A}(q^{-1})C(q^{-1})}{A(q^{-1})\hat{C}(q^{-1})} e(t)\right]^2$$

and g(x) is the gradient of

$$\mathcal{E}\left[\frac{\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1})}{A(q^{-1})\hat{C}(q^{-1})} u(t)\right]^2$$

where the expressions for $\hat{b}_i$ are used.

The quantity $\varepsilon = 1/\lambda^2$ is proportional to S.

Since (according to assumptions i) and iii)) $f(x) = 0$ has a unique solution given by

$$\hat{A}(q^{-1}) \equiv A(q^{-1}) \qquad \hat{C}(q^{-1}) \equiv C(q^{-1})$$

and f' is non singular in this point it follows from Lemma 2.8 that (5.8) has a unique solution in Ω. Since $\hat{\theta} = \theta$ is a local minimum point it follows that it is the only local minimum point of V in Ω.

Q.E.D.

Discussion of Assumptions and Results.

The assumptions i) - iii) of Theorem 5.1 are sufficient (and almost necessary) conditions for a unique global minimum, Lemma 3.5.

The assumptions i) - iii) of Theorem 5.2 are slightly stronger than the conditions used in Lemma 3.5.

If assumption i) in Theorem 5.1 is changed to $n_\ell > 0$ the mathematical machinery of Söderström (1972) will give that every local minimum points are close to some global minimum point. It is harder to examine if there are local minimum points which are not global minimum points. Since the case $n_\ell > 0$ is rather degenerated it is the author's point of view that a careful analysis is of little interest.

The very strong assumptions in the theorems are the restrictions of the signal to noise ratio. It is shown that a sufficiently high and a sufficiently small signal to noise ratio will imply existence of a unique local minimum point. However, it is unfortunately not practically possible to give any estimates of the bounds S_0 and S_1.

VI. LOCAL MINIMUM POINTS FOR STRUCTURE 6.

The structure is given by

$$y(t) = \frac{B(q^{-1})}{A(q^{-1})}\, u(t) + \frac{C(q^{-1})}{D(q^{-1})}\, e(t)$$

and the loss function for this structure can be written

$$2W(\hat{\theta}) = W_1(\hat{\theta}) + W_2(\hat{\theta})$$

$$W_1(\hat{\theta}) = \mathcal{E}\left[\left\{\frac{\hat{B}(q^{-1})}{\hat{A}(q^{-1})} - \frac{B(q^{-1})}{A(q^{-1})}\right\} \frac{\hat{D}(q^{-1})}{\hat{C}(q^{-1})}\, u(t)\right]^2 \qquad (6.1)$$

$$W_2(\hat{\theta}) = E\left[\frac{\hat{D}(q^{-1})C(q^{-1})}{\hat{C}(q^{-1})D(q^{-1})}\, e(t)\right]^2$$

If the operator $\hat{D}(q^{-1})/\hat{C}(q^{-1})$ has no influence on the number of stationary points of $W_1(\hat{\theta})$ the properties of this function is already known from Chapter 4. The function $W_2(\hat{\theta})$ is exactly the loss function for structure 3. In order to utilize these facts the following condition is introduced.

<u>Definition 6.1</u>. The function

$$\tilde{W}(\hat{\theta}) = \mathcal{E}\left[\left(\frac{\hat{B}(q^{-1})}{\hat{A}(q^{-1})} - \frac{B(q^{-1})}{A(q^{-1})}\right) u(t)\right]^2 \qquad (6.2)$$

is said to fulfil the uniqueness condition (abbreviated UC) if for u(t) persistently exciting of order $\max(\hat{n}_a+n_b,\ n_a+\hat{n}_b)$ it follows that all local minimum points satisfy

$$\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1}) \equiv 0 \tag{6.3}$$

From Chapter 4 it is known that UC holds at least in the case $\hat{n}_a = n_a = 1$.

Theorem 6.1. Consider the loss function

$$2W(\hat{\theta}) = \mathcal{E}\left[\left(\frac{\hat{B}(q^{-1})}{\hat{A}(q^{-1})} - \frac{B(q^{-1})}{A(q^{-1})}\right)\frac{\hat{D}(q^{-1})}{\hat{C}(q^{-1})}\, u(t)\right]^2 +$$

$$+ E\left[\frac{\hat{D}(q^{-1})C(q^{-1})}{\hat{C}(q^{-1})D(q^{-1})}\, e(t)\right]^2$$

Assume that

i) $\hat{n}_a \geqslant n_a$, $\hat{n}_b \geqslant n_b$, $\hat{n}_c \geqslant n_c$, $\hat{n}_d \geqslant n_d$

ii) $A(q^{-1})$ and $B(q^{-1})$ as well as $C(q^{-1})$ and $D(q^{-1})$ are relatively prime

iii) $u(t)$ is persistently exciting of order $\max(\hat{n}_a+n_b, n_a+\hat{n}_b)$

iv) The UC is fulfilled.

Then all local minimum points of $W(\hat{\theta})$ are global minimum points, i.e. they fulfil

$$\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1}) \equiv 0$$
$$\hat{C}(q^{-1})D(q^{-1}) - C(q^{-1})\hat{D}(q^{-1}) \equiv 0$$

Proof. Let θ^* be a local minimum point of $W(\hat{\theta})$. Then there is a $\delta > 0$ such that $||\theta^*-\hat{\theta}|| < \delta$ implies $W(\theta^*) \leq$ $\leq W(\hat{\theta})$. Let especially $\hat{\theta}$ coincide with θ^* in the $\hat{c}_i$- and $\hat{d}_i$-components. Then $W_1(\theta^*) \leq W_1(\hat{\theta})$. Thus θ^* is also a local minimum point of $W_1(\hat{\theta})$. From UC it follows that

$$\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1}) \equiv 0$$

When this expression is used in $W_{\hat{\theta}} = 0$ it follows that it is necessary that $(\hat{c}_1, \ldots, \hat{c}_{n_c}, \hat{d}_1, \ldots \hat{d}_{n_d})$ is a stationary point of $W_2(\hat{\theta})$, i.e.

$$\hat{C}(q^{-1})D(q^{-1}) - C(q^{-1})\hat{D}(q^{-1}) \equiv 0$$

Q.E.D.

Corr. If especially $\min(\hat{n}_a-n_a, \hat{n}_b-n_b) = 0$ and $\min(\hat{n}_c-n_c, \hat{n}_d-n_d) = 0$ the loss function has a unique local minimum point $\hat{\theta} = \theta$.

The conditions in the theorem for a unique local minimum point are partly the same conditions as used in Lemma 2.6 for a unique global minimum point, partly the uniqueness condition (UC). In contrast to the theorems for structure 5 no assumptions on the signal to noise ratio have been done.

VII. ACKNOWLEDGEMENTS.

The author wants to thank Prof. K.J. Åström for suggesting the subject and his valuable guidance throughout the work.

He also wants to thank tekn.lic. Ivar Gustavsson for several interesting discussions.

It is a great pleasure to thank Mrs. Gudrun Christensen, who typed the manuscript, and Mr. Bengt Lander, who drew the figures.

The project was partially supported by the Swedish Board for Technical Development under contract 72-202/U137.

VIII. REFERENCES.

Åström, K.J. (1968).
Lectures on the Identification Problem - the Least Squares Method. Report 6806, Division of Automatic Control, Lund Institute of Technology.

Åström, K.J. (1970).
Introduction to Stochastic Control Theory. Academic Press.

Åström, K.J. - Bohlin, T. (1966).
Numerical Identification of Linear Dynamic Systems from Normal Operating Records. Paper, IFAC Symposium on Theory of Self-Adaptive Systems, Teddington, England. In Theory of Self-Adaptive Control Systems (Ed. P.H. Hammond), Plenum Press, New York.

Åström, K.J. - Söderström, T. (1973).
Uniqueness of the Maximum Likelihood Estimates of the Parameters of a Mixed Autoregressive Moving Average Process. Report 7306, Division of Automatic Control, Lund Institute of Technology.

Bellman, R. - Åström, K.J. (1970).
On Structural Identifiability. Mathematical Bioscience, 7, 329-339.

Bohlin, T. (1970).
On the Maximum Likelihood Method of Identification. IBM J. Res. and Dev., 14, No. 1, 41-51.

Bohlin, T. (1971).
On the Problem of Ambiguities in Maximum Likelihood Identification. Automatica 7, 199-210.

Clarke, D.W. (1967).
Generalized Least Squares Estimation of the Parameters of a Dynamic Model. 1st IFAC Symposium on Identification in Automatic Control Systems, Prague.

Dickson, L.E. (1922).
First Course in the Theory of Equations, Wiley.

Kaufman, I. (1969).
The Inversion of the Vandermonde Matrix and the Transformation to the Jordan Cononical Form. IEEE Trans. Aut. Control, AC - 14, p. 774 - 777.

Ljung, L. (1971).
Characterization of the Concept of "Persistently Exciting" in the Frequence Domain. Report 7119, Division of Automatic Control, Lund Institute of Technology.

Söderström, T. (1972).
On the Convergence Properties of the Generalized Least Squares Identification Method. Report 7228, Division of Automatic Control, Lund Institute of Technology.

APPENDIX A.

This appendix deals with the degenerated solutions of (4.2). If $\hat{B}(q^{-1}) \equiv 0$ the condition $W'_{\hat{\theta}}(\hat{\theta}) = 0$ gives

$$\mathcal{E}\left[\frac{B(q^{-1})}{A(q^{-1})}\,u(t)\right]\left[\frac{q^{-i}}{\hat{A}(q^{-1})}\,u(t)\right] = 0 \qquad 1 \leqslant i \leqslant \hat{n}_b \qquad (4.8)$$

Consider for a while the following example. Let $B(q^{-1}) = bq^{-1} \not\equiv 0$, $n_a = \hat{n}_a = 2$. Further the input $u(t)$ is assumed to fulfil $u(t) = A(q^{-1})(1+c_1q^{-1}+c_2q^{-2})w(t)$ where $w(t)$ is white noise.

The equation (4.8) gives after a simple calculation

$$(1+a_1c_1+c_1^2+a_1c_1c_2+a_2c_2+c_2^2) + (-a_1c_2-c_1-c_1c_2)\hat{a}_1 +$$

$$+ c_2\hat{a}_1^2 - c_2\hat{a}_2 = 0 \qquad (A.1)$$

For $c_2 \not\equiv 0$ (A.1) describes a parabola in the $(\hat{a}_1,\hat{a}_2)$-plane. Let S be the subset of the $(\hat{a}_1,\hat{a}_2)$-plane such that $(\hat{a}_1,\hat{a}_2) \in S$ implies that the zeros of $1 + \hat{a}_1z + \hat{a}_2z^2 = 0$ are outside the unit circle. Depending on the values of a_1, a_2, c_1 and c_2 the parabola may intersect the set S.

In Figure A.1 the parabola and the set S are drawn for the special case $a_1 = -1.8$, $a_2 = 0.81$, $c_1 = 1.8$, $c_2 = 0.81$.

The following discussion will show that all stationary points, which satisfy (4.8), are saddle points.

Let θ^* satisfy $\hat{B}(q^{-1}) \equiv 0$ and (4.8). The matrix formed by

$$W_{\hat{b}_i\hat{b}_j} = \mathcal{E}\left[\frac{1}{\hat{A}(q^{-1})}\,u_{t-i}\right]\left[\frac{1}{\hat{A}(q^{-1})}\,u_{t-j}\right]$$

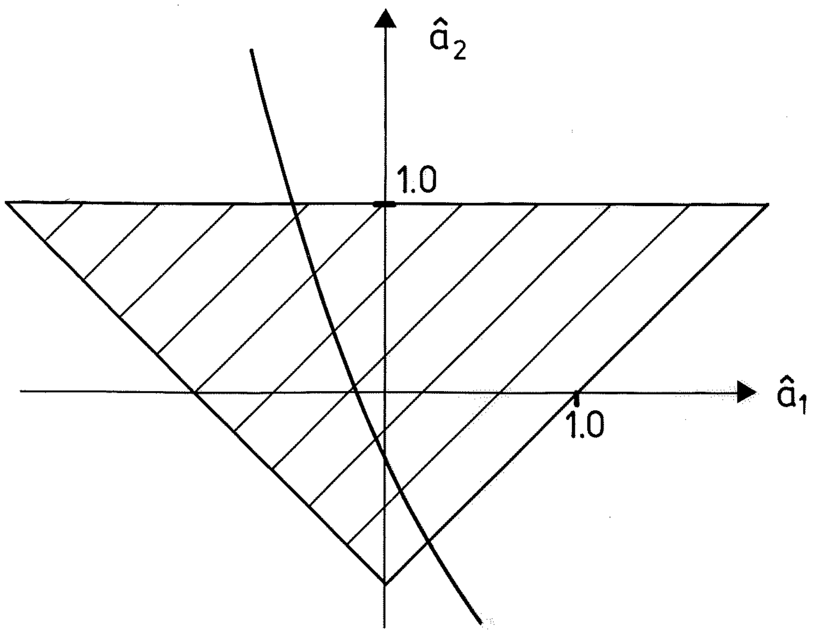


Figure A.1 - Illustration of eq. (A.1).

is positive definite for all arguments $\hat{\theta}$. If only the $\hat{b}_i$-components of θ^* are changed $W(\hat{\theta})$ will increase. If only the $\hat{a}_i$-components of θ^* are changed $W(\hat{\theta})$ will have the same value. There exists a point θ^{**} which

1) is arbitrary close to θ^*

2) differs from θ^* only in the $\hat{a}_i$-components

3) does not satisfy (4.8).

Clearly from 2) $W(\theta^{**}) = W(\theta^*)$. Given θ^{**} a new point θ^{***} is constructed. $W(\hat{\theta})$ is minimized with respect to the $\hat{b}_i$-parameters and with the $\hat{a}_i$-parameters given by θ^{**}. Since $W_{\hat{b}_i\hat{b}_j}$ is positive definite the optimization problem has a well defined solution. Call it θ^{***}. According to 3) $W(\theta^{***}) < W(\theta^{**}) = W(\theta^*)$. Finally it is observed that $||\theta^{***}-\theta^{**}||$ depends continuously on $||\theta^{**}-\theta^*||$. To summarize this means that there exists a point θ^{***} arbitrary close to θ^*, such that $W(\theta^{***}) <$ $< W(\theta^*)$. This discussion proves that θ^* must be a saddle point.

The following schematic figures are intended as an explanation of the behaviour of $W(\hat{\theta})$.

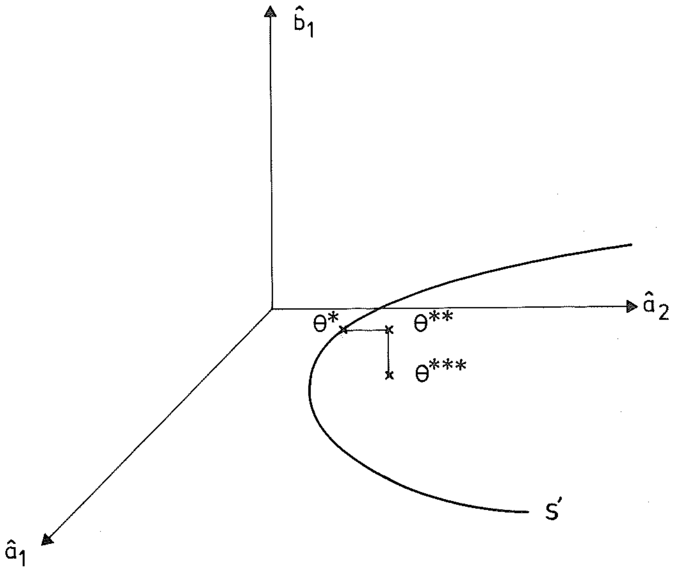


Figure A.2 - Schematic figure of θ^*, θ^{**}, θ^{***}.

The curve S' is given by (4.8) and lies in the $(\hat{a}_1,\hat{a}_2)$-plane. θ^* lies on S'. θ^{**} lies in the $(\hat{a}_1,\hat{a}_2)$-plane, close to θ^* but not on S'. θ^{***} lies below the $(\hat{a}_1,\hat{a}_2)$-plane. In Figure A.3 it is shown how $W(\hat{\theta})$ may vary in the plane spanned by θ^*, θ^{**}, θ^{***}.

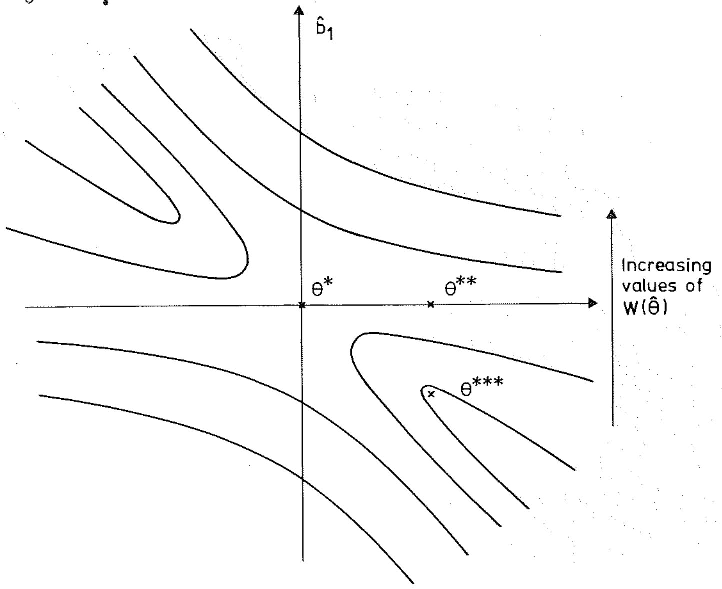


Figure A.3 - Schematic curves of $W(\hat{\theta})$ = constant.

APPENDIX B

In this appendix the structure 4 will be considered in the special case when the input signal is white noise. First the rank of the matrix Q_o defined in (4.7) will be examined. Then the equation (4.8) will be discussed.

In order to simplify the analysis it is assumed that

$$\hat{n} = \hat{n}_a = \hat{n}_b, \qquad n = n_a = n_b, \quad \hat{n} \geq n \tag{B.1}$$

This is a mild condition. Further let u(t) be white noise of unit variance. Denote $\bar{n}_a$ and $\bar{n}_b$ by $\bar{n}$, which particularly means $m = n+\bar{n}$. Define

$$A^*(z) = z^n A(z^{-1}) = z^n + \sum_{i=1}^{n} a_i z^{n-i}$$

$$\bar{A}^*(z) = z^{\bar{n}} \bar{A}(z^{-1}) = z^{\bar{n}} + \sum_{i=1}^{\bar{n}} \bar{a}_i z^{\bar{n}-i}$$

Then the ij:th element of Q_o can be written as

$$Q_{o,ij} = \frac{1}{2\Pi i} \oint \frac{z^i}{\hat{A}(z)\bar{A}(z)} \frac{z^{-j} z^{n+\bar{n}}}{A^*(z)\bar{A}^*(z)} \frac{dz}{z} \qquad \begin{matrix} i = 1,\ldots \bar{n}+\hat{n} \\ j = 1,\ldots \bar{n}+n \end{matrix} \tag{B.2}$$

The matrix Q_o will be factorized using ideas from Åström--Söderström (1973).

The poles inside the unit circle of (B.2) are exactly the zeros of $A^*(z)\bar{A}^*(z)$. They are relabelled by

$$A^*(z)\bar{A}^*(z) = \prod_{k=1}^{p} (z-u_k)^{t_k} \tag{B.3}$$

where $t_k \geq 1$, $u_k \neq u_\ell$ if $k \neq \ell$ and

$$\sum_{k=1}^{p} t_k = n + \bar{n} = m$$

With the use of (B.3) $Q_{o,ij}$ is evaluated as follows.

$$Q_{o,ij} = \frac{1}{2\Pi i} \oint \frac{z^{m+i-j-1}}{\hat{A}(z)\bar{A}(z)} \frac{1}{\prod_{k=1}^{p} (z-u_k)^{t_k}} dz$$

$$= \sum_{\ell=1}^{p} \operatorname*{Res}_{z=u_\ell} \frac{z^{m+i-j-1}}{\hat{A}(z)\bar{A}(z)} \frac{1}{\prod_{k=1}^{p} (z-u_k)^{t_k}}$$

$$= \sum_{\ell=1}^{p} \frac{1}{(t_\ell-1)!} D^{(t_\ell-1)} [z^{m+i-j-1} F_\ell(z)]_{z=u_\ell}$$

$$= \sum_{\ell=1}^{p} \sum_{k=0}^{t_\ell-1} \frac{1}{k!(t_\ell-1-k)!} D^{(k)}[z^{i-1}]_{z=u_\ell} D^{(t_\ell-1-k)} [z^{m-j} F_\ell(z)]_{z=u_\ell} \tag{B.4}$$

where D denotes differentiation with respect to z and the functions $F_\ell(z)$ are defined by

$$F_\ell(z) = \frac{1}{\hat{A}(z)\bar{A}(z)} \frac{1}{\prod\limits_{\substack{k=1\\k\neq\ell}}^{p} (z-u_k)^{t_k}} \tag{B.5}$$

Thus

$$Q_o = V\cdot\tilde{Q} = [V_1 \; V_2 \cdots \; V_p] \begin{bmatrix} \tilde{Q}_1 \\ \tilde{Q}_2 \\ \cdot \\ \cdot \\ \cdot \\ \tilde{Q}_p \end{bmatrix} \tag{B.6}$$

where $V_\ell (1\leq \ell \leq p)$ is the $(\hat{n}+\bar{n})\times t_\ell$ matrix

$$V_\ell = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ z & 1 & & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ z^{\hat{n}+\bar{n}-1} & D[z^{\hat{n}+\bar{n}-1}] & \cdots & D^{(t_\ell-1)}[z^{\hat{n}+\bar{n}-1}] \end{bmatrix}_{z=u_\ell} \tag{B.7}$$

The matrix $\tilde{Q}_\ell (1\leq\ell\leq p)$ is $t_\ell\times m$ and is given by

$$\tilde{Q}_{\ell,ij} = \frac{1}{(i-1)!(t_\ell-i)!} D^{(t_\ell-i)}[z^{m-j}F_\ell(z)]_{z=u_\ell} \tag{B.8}$$

The matrix V is a generalization of the van der Monde matrix. It follows from Kaufamn (1969) that the rank of V is m.

The matrix $\tilde{Q}$ also can be factorized. In fact

$$\tilde{Q} = S \cdot X \tag{B.9}$$

where X is an m×m matrix which can be written as

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$$

The matrix $X_\ell (1 \leq \ell \leq p)$ is $t_\ell \times m$ and holds

$$X_\ell = \begin{bmatrix} z^{m-1} & \cdots & 1 \\ D[z^{m-1}] & & 0 \\ \cdot & & \cdot \\ \cdot & & \cdot \\ D^{(t_\ell -1)}[z^{m-1}] & \cdots & 0 \end{bmatrix}_{z=u_\ell} \tag{B.10}$$

According to Kaufman (1969) X is nonsingular. The square matrix S can be written as

$$S = \begin{bmatrix} S_1 & & & & \\ & S_2 & & 0 & \\ & & \ddots & & \\ & 0 & & & \\ & & & & S_p \end{bmatrix}$$

where $S_1, \ldots, S_p$ are square block matrices of the orders $t_1 \times t_1, \ldots, t_p \times t_p$. They are given by

$$S_{\ell,ik} = \begin{cases} 0 \text{ if } \quad k > t_\ell + 1 - i \\ \dfrac{1}{(i-1)!(k-1)!(t_\ell - i+1-k)!} D^{(t_\ell - i+1-k)} [F_\ell(z)]_{z=u_\ell} \\ \qquad \text{if} \quad k \leq t_\ell + 1 - i \end{cases}$$

This means that S_ℓ has the following structure

$$S_\ell = \begin{bmatrix} S_{\ell,11} & \cdots & S_{\ell,1t_\ell} \\ \vdots & \iddots & \\ S_{\ell,t_\ell 1} & & 0 \end{bmatrix}$$

The elements of the cross diagonal are given by

$$S_{\ell,i\ t_\ell+1-i} = \frac{1}{(i-1)!(t_\ell - i)!} F_\ell(u_\ell) \qquad 1 \leq i \leq t_\ell$$

and they are nonzero according to the definition (B.5) of $F_\ell(z)$. This means that S is nonsingular.

Thus it has been proven that the rank of Q_o is m.

Now the degenerated case of $\hat{B}(q^{-1}) \equiv 0$ will be treated. Consider the equation (4.8). Factorize $A^*(z)$ as

$$A^*(z) = \prod_{i=1}^{q} (z-u_j)^{s_j} \tag{B.11}$$

where $s_j \geq 1$ and

$$\sum_{j=1}^{q} s_j = n$$

Using (B.11) the equation (4.8) is written as

$$\frac{1}{2\Pi i} \oint \frac{B^*(z)}{\prod\limits_{j=1}^{q} (z-u_j)^{s_j}} \, \frac{z^i}{\hat{A}(z)} \, dz = 0 \qquad 0 \leq i \leq \hat{n} -1 \tag{B.12}$$

Straight-forward calculations analogous to (B.4) give

$$U \cdot g = [U_1 \; U_2 \; \ldots U_q] \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_q \end{bmatrix} = 0 \tag{B.13}$$

The matrix $U_\ell (1 \leq \ell \leq q)$ is $\hat{n} \times s_\ell$ and holds

$$U_\ell = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ z & 1 & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ z^{\hat{n}-1} & D[z^{\hat{n}-1}] & \ldots & D^{(s_\ell -1)}[z^{\hat{n}-1}] \end{bmatrix}_{z=u_\ell} \tag{B.14}$$

To summarize it has been shown that if the input is white noise and the assumptions (B.1) hold, then the loss function has no other stationary points than the global minimum points.

Is it possible to extend the calculations? One extension would be to substitute (B.1) with the more general $\min(\hat{n}_a - n_a,\ \hat{n}_b - n_b) \geq 0$ and another to permit input signals which are filtered white noise. Note that it is trivial to allow the case $\hat{n} \geq \max(n_a, n_b)$. If e.g. n_a is larger than n_b the polynomial $B^*(z)$ can be multiplied by $z^{n_a - n_b}$ and the new polynomial will have n_a coefficients.

However, the extensions desired are not possible in general. The reason is that the number of poles inside the unit circle may be larger than the numbero of rows in Q_o resp. the number of equations in (B.12). This means that the matrices V and U will have a smaller number of rows than columns which causes the idea of the calculations to break down.

However, there are cases where the results can be extended further. For instance, the analysis of (4.8) can be extended in a straight-forward way to the case

$$\max(n_a, n_b) \leq \hat{n}_b, \quad \hat{n}_a \text{ arbitrary}$$

Since the extensions can not be done in general and since the assumptions (B.1) are mild, it is of minor interest to extend the calculations further.

The vector g_ℓ is given by

$$g_{\ell,i} = \frac{1}{(i-1)!(s_\ell-i)!} D^{(s_\ell-i)} [f_\ell(z)]_{z=u_\ell} \qquad (B.15)$$

The functions $f_\ell(z)$ are given by

$$f_\ell(z) = \frac{B^*(z)}{\hat{A}(z) \prod_{\substack{j=1 \\ j\neq\ell}}^{q} (z-u_j)^{s_j}} \qquad (B.16)$$

Since the rank of U is n, see Kaufman (1969), it follows that

$$g_\ell = 0 \qquad 1 \leq \ell \leq q$$

Then it follows from Åström-Söderström (1973) that

$$D^{(k)}[B^*(z)]_{z=u_j} = 0 \qquad 0 \leq k \leq s_j-1, \quad 1 \leq j \leq q \qquad (B.17)$$

Thus

$$B^*(z) = \tilde{B}(z) \prod_{j=1}^{q} (z-u_j)^{s_j}$$

where $\tilde{B}(z)$ is some polynomial. $B^*(z)$ is, however, a polynomial of degree n-1, while the product $\prod_{j=1}^{q} (z-u_j)^{s_j}$ is a polynomial of degree n. This implies that $\tilde{B}(z) \equiv 0$ and the contradiction $B(z) \equiv 0$ is established.

To summarize it has been shown that if the input is white noise and the assumptions (B.1) hold, then the loss function has no other stationary points than the global minimum points.

Is it possible to extend the calculations? One extension would be to substitute (B.1) with the more general $\min(\hat{n}_a - n_a,\ \hat{n}_b - n_b) \geq 0$ and another to permit input signals which are filtered white noise. Note that it is trivial to allow the case $\hat{n} \geq \max(n_a, n_b)$. If e.g. n_a is larger than n_b the polynomial $B^*(z)$ can be multiplied by $z^{n_a - n_b}$ and the new polynomial will have n_a coefficients.

However, the extensions desired are not possible in general. The reason is that the number of poles inside the unit circle may be larger than the numbero of rows in Q_o resp. the number of equations in (B.12). This means that the matrices V and U will have a smaller number of rows than columns which causes the idea of the calculations to break down.

However, there are cases where the results can be extended further. For instance, the analysis of (4.8) can be extended in a straight-forward way to the case

$$\max(n_a, n_b) \leq \hat{n}_b, \quad \hat{n}_a \text{ arbitrary}$$

Since the extensions can not be done in general and since the assumptions (B.1) are mild, it is of minor interest to extend the calculations further.

AN ON-LINE ALGORITHM FOR APPROXIMATE MAXIMUM LIKELIHOOD IDENTIFICATION OF LINEAR DYNAMIC SYSTEMS

T. SÖDERSTRÖM

Report 7308, March 1973
Lund Institute of Technology
Division of Automatic Control

TABLE OF CONTENTS

AN ON-LINE ALGORITHM FOR APPROXIMATE MAXIMUM LIKELIHOOD IDENTIFICATION OF LINEAR DYNAMIC SYSTEMS

T. Söderström

ABSTRACT

A recursive algorithm for maximum likelihood estimation of parameters in a linear dynamic system is presented. The basic idea in the algorithm is a recursive optimization of the likelihood function. Different approximations are used. With special simplifications the algorithm becomes identical to methods earlier proposed. The properties of the algorithm are illustrated by application to data from simulated systems as well as plant measurements.

I. INTRODUCTION

In the field of the identification of dynamic systems special interest has been given to on-line methods. It may be desirable to proceed the identification until a specified accuracy is achieved. An on-line identification method also is necessary for adaptive control.

Several on-line identification methods have been proposed. In Åström-Eykhoff (1971) a short description of different methods is given. The algorithms described in Young (1970), Young-Shellswell-Nethling (1971) seem to work quite satisfactorily.

When off-line methods are considered it is known that the maximum likelihood method is a powerful one and in most cases gives the "best" estimates, Åström-Bohlin (1966), Gustavsson (1969b). The purpose of this report is to describe an approximative recursive version of this method using ideas due to Åström, who has made an outline of the algorithm.

It is well-known, Åström (1968), Åström-Eykhoff (1971) that the least squares (LS) method easily can be computed recursivly. The recursive version can be interpreted as a Kalman filter. The ML method can be considered as an extension of the LS method. One way to construct a recursive ML algorithm is to generalize the Kalman filter of the LS case. This approach has been taken by Young for the estimation of parameters of time series.

Panuška (1968) gives a similar algorithm based on stochastic approximation. A comparison of Panuška's algorithm and the off-line ML method is given in Valis-Gustavsson (1969).

In this report an estimation algorithm will be derived via a recursive minimization of a time varying loss function. When different approximations and simplifications are made the algorithm is the same as the one used by Young or the one used by Panuška.

The approach of minimizing a loss function can be applied to different models. Several well-known methods as least squares, generalized least squares, Clarke (1967), Söderström (1972), the "ordinary" ML, Åström-Bohlin (1966) and the method used by Bohlin (1970) can be interpreted as maximum likelihood models when appropriate assumptions of the structure of the systems are made. All the methods can be expressed as a minimization of a loss function of the form

$$V_N(\hat{\theta}) = \frac{1}{2} \sum_{t=1}^{N} \varepsilon^2(t;\hat{\theta}) \tag{1.1}$$

N is the number of samples and $\varepsilon(t;\hat{\theta})$ the residual at time t. The vector $\hat{\theta}$ is an estimate of θ, a vector containing parameters which describe the system. The elements of $\hat{\theta}$ will be called the model parameters. The explicit expression of $\varepsilon(t;\hat{\theta})$ as a function of $\hat{\theta}$ differs between the different methods. The variances of the residuals can be estimated by

$$\hat{\lambda}^2 = \frac{2}{N} \min_{\hat{\theta}} V_N(\hat{\theta}) \tag{1.2}$$

Let $\hat{\theta}_N$ minimize $V_N(\hat{\theta})$. A recursive algorithm must give $\hat{\theta}_{N+1}$ from $\hat{\theta}_N$, the measurements at time N+1 and a reasonably small amount of collected information of the system. In the recursive LS method this is done exactly but for the other methods approximations have to be used.

Another way of discussing the properties of a reasonable algorithm it to use the concept of sufficient statistics. When the disturbances are gaussian it is well-known that there is a sufficient statistic in the LS case, namely

$V''_N(\hat{\theta}_{N-1})$, $\hat{\theta}_{N-1}$ and a few of the latest measurements. In the general case a sufficient statistic must include all old measurements explicitly. Thus it is suitable to base an algorithm on an approximate sufficient statistic.

In the next chapter an algorithm for the recursive minimization of V_N is developed. Different approximations are discussed. In chapter III the Kalman filter approach is taken into consideration and some comparisons are made. Possible limits to which the estimates may converge are analysed in chapter IV. The fifth chapter contains some examples and discussions about how to implement the algorithm. Finally examples using plant measurements are presented in chapter VI.

II. A RECURSIVE MAXIMUM LIKELIHOOD ESTIMATOR

In this chapter the recursive algorithm is developed. The first part deals with the recursive minimization of the loss function

$$V_N(\hat{\theta}) = \frac{1}{2} \sum_{t=1}^{N} \varepsilon^2(t;\hat{\theta}) \qquad (1.1)$$

in general. In the second part the algorithm will be applied to the specific model, Åström-Bohlin (1966)

$$\hat{A}(q^{-1})y(t) = \hat{B}(q^{-1})u(t) + \hat{C}(q^{-1})\varepsilon(t;\hat{\theta}) \qquad (2.1)$$

where y(t) is the output and u(t) the input at time t. The polynomial operators are

$$\hat{A}(q^{-1}) = 1 + \hat{a}_1 q^{-1} + \ldots + \hat{a}_n q^{-n}$$

$$\hat{B}(q^{-1}) = \hat{b}_1 q^{-1} + \ldots + \hat{b}_n q^{-n}$$

$$\hat{C}(q^{-1}) = 1 + \hat{c}_1 q^{-1} + \ldots + \hat{c}_n q^{-n}$$

q^{-1} is the backward shift operator and

$$\hat{\theta} = [\hat{a}_1 \ldots \hat{a}_n \hat{b}_1 \ldots \hat{b}_n \hat{c}_1 \ldots \hat{c}_n]^T$$

Let $\hat{\theta}_N$ be the minimum point of $V_N(\hat{\theta})$. The estimate $\hat{\theta}_{N+1}$ will be computed from a Taylor expansion of $V_{N+1}(\hat{\theta})$ around θ_N. Suppose that an expansion including second order terms is accurate enough.

$$V_{N+1}(\hat{\theta}) \approx V_{N+1}(\hat{\theta}_N) + V'_{N+1}(\hat{\theta}_N)(\hat{\theta}-\hat{\theta}_N) +$$

$$+ \frac{1}{2}(\hat{\theta}-\hat{\theta}_N)^T V''_{N+1}(\hat{\theta}_N)(\hat{\theta}-\hat{\theta}_N)$$

Minimization gives

$$\hat{\theta}_{N+1} = \hat{\theta}_N - V''_{N+1}(\hat{\theta}_N)^{-1} V'_{N+1}(\hat{\theta}_N)^T \qquad (2.2)$$

which is the first iteration of a Newton Raphson algorithm applied to the equation $V'_{N+1}(\hat{\theta}) = 0$.

The estimated minimum value of $V_{N+1}(\hat{\theta})$ is

$$V_{N+1}(\hat{\theta}_{N+1}) = V_{N+1}(\hat{\theta}_N) - \frac{1}{2} V'_{N+1}(\hat{\theta}_N) V''_{N+1}(\hat{\theta}_N)^{-1} V'_{N+1}(\hat{\theta}_N)^T \qquad (2.3)$$

To form a recursive estimator the relation between $V_{N+1}(\hat{\theta})$ and $V_N(\hat{\theta})$ must be utilized. By definition

$$V_{N+1}(\hat{\theta}) = V_N(\hat{\theta}) + \frac{1}{2}\varepsilon^2(N+1;\hat{\theta}) \qquad (2.4)$$

$$V'_{N+1}(\hat{\theta}) = V'_N(\hat{\theta}) + \varepsilon(N+1;\hat{\theta})\varepsilon'(N+1;\hat{\theta}) \qquad (2.5)$$

$$V''_{N+1}(\hat{\theta}) = V''_N(\hat{\theta}) + \varepsilon'(N+1;\hat{\theta})^T \varepsilon'(N+1;\hat{\theta}) + \varepsilon(N+1;\hat{\theta})\varepsilon''(N+1;\hat{\theta}) \qquad (2.6)$$

The following approximations are made now

$$V'_N(\hat{\theta}_N) = 0 \qquad (2.7)$$

$$\varepsilon(N+1;\hat{\theta}_N)\varepsilon''(N+1;\hat{\theta}_N) = 0 \qquad (2.8)$$

$$V''_N(\hat{\theta}_N) = V''_N(\hat{\theta}_{N-1}) \qquad (2.9)$$

The assumption (2.7) can be assumed to hold since $\hat{\theta}_N$ is assumed to minimize $V_N(\hat{\theta})$. For off-line ML the term $\sum_{t=1}^{N} \varepsilon(t;\hat{\theta})\varepsilon''(t;\hat{\theta})$ has little influence on the minimization, Gustavsson (1969b). The equation (2.9) is motivated if $\hat{\theta}_N$ is close to $\hat{\theta}_{N-1}$. Also notice that (2.7) - (2.9) as well as the Taylor expansion hold exactly in the LS case.

With the use of the approximations

$$V_{N+1}(\hat{\theta}_{N+1})=V_N(\hat{\theta}_N)+\frac{1}{2}\varepsilon^2(N+1;\hat{\theta}_N)-\frac{1}{2}V'_{N+1}(\hat{\theta}_N)V''_{N+1}(\hat{\theta}_N)^{-1}V'_{N+1}(\hat{\theta}_N)^T \tag{2.10}$$

$$V'_{N+1}(\hat{\theta}_N) = \varepsilon(N+1;\hat{\theta}_N)\,\varepsilon'(N+1;\hat{\theta}_N) \tag{2.11}$$

$$V''_{N+1}(\hat{\theta}_N) = V''_N(\hat{\theta}_{N-1})+\varepsilon'(N+1;\hat{\theta}_N)^T\varepsilon'(N+1;\hat{\theta}_N) \tag{2.12}$$

Introduce the notations

$$P_N = V''_N(\hat{\theta}_{N-1})^{-1} \tag{2.13}$$

$$\varphi_N = \varepsilon'(N;\ \hat{\theta}_{N-1})^T \tag{2.14}$$

$$\varepsilon_N = \varepsilon(N;\hat{\theta}_{N-1}) \tag{2.15}$$

$$\gamma_{N+1} = 1 + \varphi_{N+1}^T P_N \varphi_{N+1} \tag{2.16}$$

Then (2.2) can be written as

$$\hat{\theta}_{N+1} = \hat{\theta}_N - P_{N+1}\varphi_{N+1}\varepsilon_{N+1} \tag{2.17}$$

The well-known matrix lemma

$$[M+bb^T]^{-1} = M^{-1} - M^{-1}b[1 + b^TM^{-1}b]^{-1}b^TM^{-1}$$

applied to (2.12) gives

$$P_{N+1} = P_N - \frac{1}{\gamma_{N+1}} P_N\varphi_{N+1}\varphi_{N+1}^T P_N \tag{2.18}$$

Finally (2.10) can be rewritten after some trivial calculations as

$$V_{N+1}(\hat{\theta}_{N+1}) = V_N(\hat{\theta}_N) + \frac{1}{2}\frac{1}{\gamma_{N+1}}\varepsilon_{N+1}^2 \tag{2.19}$$

In the general case it now remains to develop recursive equations for ε_N and φ_N. For the LS case this is very simple since $\varepsilon(t;\hat{\theta})$ is linear in $\hat{\theta}$. The derived algorithm coincides with the well-known one in the LS case. The expression (2.19) can be found in Wieslander (1971) where it is derived using a Kalman filter representation.

In the derivation of recursive equations for ε_N and φ_N specialization will be made to the model

$$\hat{A}(q^{-1})y(t) = \hat{B}(q^{-1})u(t) + \hat{C}(q^{-1})\varepsilon(t;\hat{\theta}) \qquad (2.1)$$

which can be written in state space form as

$$x(t+1) = \begin{bmatrix} -\hat{c}_1 & 1 & & 0 \\ \vdots & & \ddots & \\ \vdots & & & 1 \\ -\hat{c}_n & 0 & & 0 \end{bmatrix} x(t) + \begin{bmatrix} 1 & \hat{a}_1 & -\hat{b}_1 \\ 0 & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 0 & \hat{a}_n & -\hat{b}_n \end{bmatrix} \begin{bmatrix} y(t+1) \\ y(t) \\ u(t) \end{bmatrix} \qquad (2.20)$$

$$\varepsilon(t;\hat{\theta}) = x_1(t)$$

The derivatives $\varepsilon'(t;\hat{\theta})$ are given by

$$\hat{C}(q^{-1}) \frac{\partial\varepsilon}{\partial\hat{a}_i}(t;\hat{\theta}) = y(t-i)$$

$$\hat{C}(q^{-1}) \frac{\partial\varepsilon}{\partial\hat{b}_i}(t;\hat{\theta}) = -u(t-i) \qquad (2.21)$$

$$\hat{C}(q^{-1}) \frac{\partial\varepsilon}{\partial\hat{c}_i}(t;\hat{\theta}) = -\varepsilon(t-i;\hat{\theta})$$

A state space form of (2.21) is

$$\varepsilon'(t+1;\hat{\theta})=\begin{bmatrix} \begin{matrix}-\hat{c}_1 & \cdots & -\hat{c}_n\\ 1 & & \\ & \ddots & \\ & 1 & 0\end{matrix} & & \mathbf{0} \\ & \begin{matrix}-\hat{c}_1 & \cdots & -\hat{c}_n\\ 1 & & \\ & \ddots & \\ & 1 & 0\end{matrix} & \\ \mathbf{0} & & \begin{matrix}-\hat{c}_1 & \cdots & -\hat{c}_n\\ 1 & & \\ & \ddots & \\ & 1 & 0\end{matrix}\end{bmatrix}\varepsilon'(t;\hat{\theta})+\begin{bmatrix}y(t)\\0\\ \vdots\\0\\-u(t)\\0\\ \vdots\\0\\-\varepsilon(t;\hat{\theta})\\0\\ \vdots\\0\end{bmatrix} \tag{2.22}$$

The initial values are $x(0)=0$, $\varepsilon'(0;\hat{\theta})=0$.

In order to compute $\varepsilon(t;\hat{\theta})$ and $\varepsilon'(t;\hat{\theta})$ (2.20) and (2.22) have to be solved from $t=0$. Since ε and ε' must be computed for new arguments at every time step this means an unreasonable lot of calculations. Moreover, all old measurements must be saved. Note that no matrix multiplications have to be done explicitly. E.g. all but three components of $\varepsilon(t+1;\hat{\theta})$ can be computed by shift.

One way of reducing the computational work is the following. (2.20) and (2.22) are solved only once and with time variable matrices. When $x(t)$ and $\varepsilon'(t;\hat{\theta}_{t-1})$ are computed from $x(t-1)$ respectively $\varepsilon'(t-1;\hat{\theta}_{t-2})$ the components of $\hat{\theta}_{t-1}$ are used in the matrices. If $\hat{\theta}_t$ does not change very much with t this approximation can be assumed to be good. The resulting values of the residual will be denoted $\hat{\varepsilon}_t$.

A further simplification would be to substitute $\hat{C}(q^{-1})$ in (2.21) with 1. This does not reduce the computations very much but it has a nice interpretation which will be shown in the next chapter.

There are other possibilities to compute approximative values of the residuals. One is the following which is used by Young (1970) and Panuška (1968). The equation (2.1) can be written as

$$\varepsilon(t) = y(t) - [-y(t-1)..-y(t-n)\ u(t-1)..u(t-n)\ \varepsilon(t-1)...\varepsilon(t-n)]\hat{\theta} \quad (2.23)$$

An exact computation of $\varepsilon(t;\hat{\theta})$ requires the solution of (2.23) from t=0 with constant $\hat{\theta}$. Similarly to the method previously described $\varepsilon(t;\hat{\theta}_{t-1})$ can be approximated by

$$\varepsilon(t;\hat{\theta}_{t-1}) = y(t) -$$

$$-[-y(t-1)..-y(t-n)\ u(t-1)..u(t-n)\ \varepsilon(t-1;\hat{\theta}_{t-2})..\varepsilon(t-n;\hat{\theta}_{t-n-1}]\hat{\theta}_{t-1}$$

$$(2.24)$$

The algorithm used by Young is obtained if (2.24) is used for computations of ε_N and (2.21) with $\hat{C}(q^{-1})$ substituted by 1 for computations of φ_N.

Panuška's algorithm uses a gradient method for the minimization. In (2.17) P_N is substituted by $\frac{K}{N}$ I where K is a suitable constant. ε_N and φ_N are computed as in Young's algorithm.

The general algorithm and Young's version are compared using simulated data in chapter V. For these examples both the methods may give bad estimates if they are applied straight-forward. Suitable modifications are discussed in chapter V. Further it turns out that after these modifications both the methods seem to work well in the present simulated systems but Young's algorithm gives larger variances of the parameter estimates. For both the methods the convergence of the $\hat{A}$-and the $\hat{B}$-parameters are considerably faster than the convergence of the $\hat{C}$-parameters.

In Valis-Gustavsson (1969) a comparison is made between Panuška's method and the off-line ML method. The comparison shows not unexpectedly that the off-line ML method is superior. Especially the C-parameters seem to be difficult to estimate accurately with Panuška's method.

III. COMPARISON WITH KALMAN FILTERING

It is well-known that the recursive least squares method can be interpreted as a Kalman filter, Åström (1968), Åström-Eykhoff (1971). Using some approximations this idea can be used for the model (2.1) as well. It turns out that Young's algorithm is very "natural" from this point of view.

The system corresponding to the model (2.1) can be written as

$$\begin{aligned} \theta(t+1) &= \theta(t) \\ y(t) &= C(t)\theta(t) + e(t) \end{aligned} \tag{3.1}$$

where e(t) is white noise with variance λ^2 and

$$C(t) = [-y(t-1)\ldots-y(t-n) \quad u(t-1)\ldots u(t-n) \quad e(t-1)\ldots e(t-n)]$$

$$\theta(t) = [a_1,\ldots, a_n, b_1,\ldots,b_n, c_1,\ldots, c_n]^T$$

If C(t) were known a Kalman filter for estimating the state $\theta(t)$ would be

$$\hat{\theta}(t+1) = \hat{\theta}(t) + K(t+1)[y(t+1) - C(t+1)\hat{\theta}(t)]$$

$$K(t) = \frac{1}{\lambda^2} P(t)C(t)^T \tag{3.2}$$

$$P(t) = P(t-1)-P(t-1)C(t)^T[\lambda^2+C(t)P(t-1)C(t)^T]^{-1}C(t)P(t-1)$$

A way to overcome the difficulty of C(t) being partly unknown is to replace e(t-1)...e(t-n) in C(t) by $\varepsilon(t-1)\ldots\varepsilon(t-n)$. The residuals $\{\varepsilon(t)\}$ are defined recursively through

$$\varepsilon(t+1) = y(t+1) - C(t+1)\hat{\theta}(t)$$

The algorithm obtained is exactly Young's method.

IV. ANALYSIS

To establish convergence of the algorithm, i.e. to prove that $\hat{\theta}_k \to \theta$, $k \to \infty$ is a very hard task. The purpose of the following analysis only is to determine possible limits of $\{\hat{\theta}_k\}$.

First it is observed that the recursive algorithm given by (2.17) - (2.19) formally can be interpreted as a recursive least squares solution of the system of equations

$$\begin{bmatrix} \varphi_1^T \\ \varphi_2^T \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \hat{\theta} = \begin{bmatrix} -\varepsilon_1 + \varphi_1^T \hat{\theta}_o \\ -\varepsilon_2 + \varphi_2^T \hat{\theta}_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \qquad (4.1)$$

This is true only formally since the right-hand side involves $\hat{\theta}_o, \hat{\theta}_1, \ldots$

Assume that $\hat{\theta}_t$ tends to θ^* with probability one when the number of samples tends to infinity. Assume that θ^* corresponds to a model for which $A^*(z)$ and $C^*(z)$ have all zeros outside the unit circle.

If the initial values of the recursive least squares solution are neglected then $\hat{\theta}_N$ must fulfil the normal equations

$$\frac{1}{N}\sum_{t=1}^{N} (\varphi_t \varphi_t^T)\hat{\theta}_N = \frac{1}{N}\sum_{t=1}^{N} \varphi_t(-\varepsilon_t + \varphi_t^T \hat{\theta}_{t-1}) \qquad (4.2)$$

It is shown in the appendix that $\hat{\theta}_N$ and $\hat{\theta}_{t-1}$ asymptotically can be replaced by θ^*. Further ε_t and φ_t (asymptotically) can be replaced by $\varepsilon(t;\theta^*)$ and $\varepsilon'(t;\theta^*)$ respectively. Thus (4.2) implies

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \varepsilon(t;\theta^*)\varepsilon'(t;\theta^*) = 0 \qquad (4.3)$$

Using standard ergodic theory it is possible to show, see Söderström (1972), that (4.3) can be substituted by

$$E\ \varepsilon(t;\theta^{*})\varepsilon'(t;\theta^{*}) = 0 \tag{4.4}$$

However, (4.4) is exactly the equation for the stationary points of

$$W(\hat{\theta}) = E[\varepsilon(t;\hat{\theta})]^{2} \tag{4.5}$$

In Söderström (1973) an analysis of the number of local minimum points of $W(\hat{\theta})$ is given. It is shown that θ is always a minimum point and conditions are given which guarantee that θ is a unique local minimum point of $W(\hat{\theta})$.

Thus if these condition are fulfilled and $\hat{\theta}_t$ converges a.s. it must converge to the correct values.

V. NUMERICAL EXAMPLES

In this chapter some numerical examples will be given. It has appeared to the author by practical experience that the algorithm cannot be successfully applied in a straightforward way, but suitable tricks make it work rather well. Several tricks and modifications have been tried by the author but only the best one is used in the examples presented. At the end of the chapter a brief discussion of other tricks is given.

Two different versions of the algorithm are used in the examples. One is called RMLE1 (Recursive Maximum Likelihood Estimation, version 1) and the other RMLE2. Both the versions include the basic algorithm given by (2.17) - - (2.19). The estimate of λ is taken as

$$\hat{\lambda} = \sqrt{\frac{2}{N} V_N(\hat{\theta}_N)}$$

In RMLE1 the residuals are computed from (2.22). RMLE2 is the version used by Young (who calls it AML, Approximate Maximum Likelihood).

The initial values of all variables involved were all chosen as 0 with the exception of P_0 which was chosen $100 \cdot I$. In the off-line version of the ML algorithm a test of stability of $\hat{C}(z)$ is made at every iteration and the estimates are modified to give stability, Gustavsson (1969b) This trick was tried in the recursive algorithm as well and it improved the result.

It would be valuable to have one number giving the accuracy of the result. For instance, one can use $||\hat{\theta}-\theta||^2$ or more generally $(\hat{\theta}-\theta)^T Q(\hat{\theta}-\theta)$ where Q is some symmetric positive definite matrix.

In the following examples an asymptotic loss function was used, namely,

$$W(\hat{\theta};\theta) = \frac{1}{\lambda^2} E\, \varepsilon^2(t)$$

where

$$\hat{C}(q^{-1})\varepsilon(t) = \hat{A}(q^{-1})y(t) - \hat{B}(q^{-1})u(t)$$

and the process is described by

$$A(q^{-1})y(t) = B(q^{-1})u(t) + C(q^{-1})e(t), \quad E\, e^2(t) = \lambda^2$$

with $\{e(t)\}$ white noise. Thus

$$W(\hat{\theta};\theta) = \frac{1}{\lambda^2} E \left[\frac{\hat{A}(q^{-1})B(q^{-1}) - A(q^{-1})\hat{B}(q^{-1})}{A(q^{-1})\hat{C}(q^{-1})} u(t) + \right.$$

$$\left. + \frac{\hat{A}(q^{-1})C(q^{-1})}{A(q^{-1})\hat{C}(q^{-1})} e(t) \right]^2$$

Assume that the input is independent of the noise. If the spectral density of the input is known (in the examples the input is treated as white noise) $W(\hat{\theta};\theta)$ can easily be computed from integrals.

Clearly, Åström-Söderström (1973), $W(\hat{\theta};\theta) \geq 1$ for all $\hat{\theta}$ where equality implies $\hat{\theta} = \theta$. Further $W_{\hat{\theta}}(\theta;\theta) = 0$.

An expected asymptotic value of $W(\hat{\theta};\theta)$ can be calculated. Assume that $\hat{\theta}$ is asymptotically gaussian distributed with mean θ and variance equal to the Cramér-Rao lower bound, i.e.

$$P_{\hat{\theta}} = \frac{2}{N} W(\theta;\theta)\, W''_{\hat{\theta}\hat{\theta}}(\theta;\theta)^{-1}$$

This assumption is valid for the off-line ML estimates, Åström-Bohlin (1966). For large values of $N, W(\hat{\theta};\theta)$ then can be approximated by $1 + \frac{1}{N} x$ where

$$x = (\hat{\theta}-\theta)^T P_{\hat{\theta}}^{-1}(\hat{\theta}-\theta)$$

is asymptotically $\chi^2(3n)$ distributed. Especially, under these assumptions $E\ W(\hat{\theta};\theta) = 1 + \frac{3n}{N}$.

In order to analyse the properties of the methods, the algorithms were applied to data from simulated systems. A number of realizations was used. The average values and the RMS errors of $\hat{\theta}$ were computed. The RMS errors are

$$\left(\frac{1}{k}\sum_{j=1}^{k}(\hat{\theta}_i(j) - \theta_i)^2\right)^{1/2}$$

where $\hat{\theta}_i(j)$ denotes the i:th component of $\hat{\theta}$ obtained at the identification of the j:th realization. The average values and the RMS errors are compared with their theoretically expected values based on the Cramér-Rao lower bound.

In all examples the number of samples was 2 000. The input signal was a PRBS with amplitude 1.0. 11 different realizations were used.

For the first order system the algorithms applied straightforward work rather satisfactorily. RMLE1 produces a considerably lower variance of $\hat{c}_1$ than RMLE2. The results are given in table 5.1.

		a_1	b_1	c_1	λ	W
Expected	mean	-0.8	1.0	0.7	1.0	1.0015
	RMS error	0.012	0.017	0.017	0.032	0.0019
RMLE1	mean	-0.796	1.005	0.695	1.009	1.0023
	RMS error	0.019	0.020	0.014	0.037	0.0028
RMLE2	mean	-0.796	1.005	0.675	1.019	1.0056
	RMS error	0.023	0.027	0.038	0.042	0.0067

Table 5.1. Results for a first order system. RMLE1 is the general algorithm given in chapter II. RMLE2 is Young's algorithm.

For a second order system, however, the results are considerably inferior than in the first order case. The results of a straight-forward application of the algorithms are given in table 5.2.

	a_1	a_2	b_1	b_2	c_1	c_2	λ	W
Expected mean	-1.5	0.7	1.0	0.5	-1.0	0.2	1.0	1.0030
RMS error	0.007	0.006	0.022	0.029	0.023	0.022	0.032	0.0032
RMLE1 mean	-1.418	0.624	0.990	0.513	-0.747	0.103	1.267	1.164
RMS error	0.259	0.242	0.073	0.072	0.517	0.116	0.591	0.185
RMLE2 mean	-1.490	0.688	1.009	0.487	-0.867	0.044	1.112	1.054
RMS error	0.029	0.027	0.028	0.072	0.180	0.161	0.140	0.076

Table 5.2. Results for a second order system. Straight-forward application of the algorithms. RMLE1 is the general algorithm given in chapter II. RMLE2 is Young's algorithm.

In Figures 5.1 and 5.2 the estimates of one of the realizations (RMLE1 is used) are plotted versus time. A comparison with table 2 shows that the result of the identification of this realization is among the best ones. From Figure 5.1 a general tendency of the algorithm can be seen. It loses its "gain" after some hundred samples and most often the estimates of the C-parameters then are not close to the correct values. This fact indicates that some kind of restarts would be valuable.

This idea will be combined with another. In the computations of the derivatives of the loss function $\hat{\varepsilon}_t$ is used as an approximation of $\varepsilon(t;\hat{\theta})$ for various values of $\hat{\theta}$. If $\hat{\theta}$ is fixed for a number of samples, the approximation $\hat{\varepsilon}_t \approx \varepsilon(t;\hat{\theta})$ probably

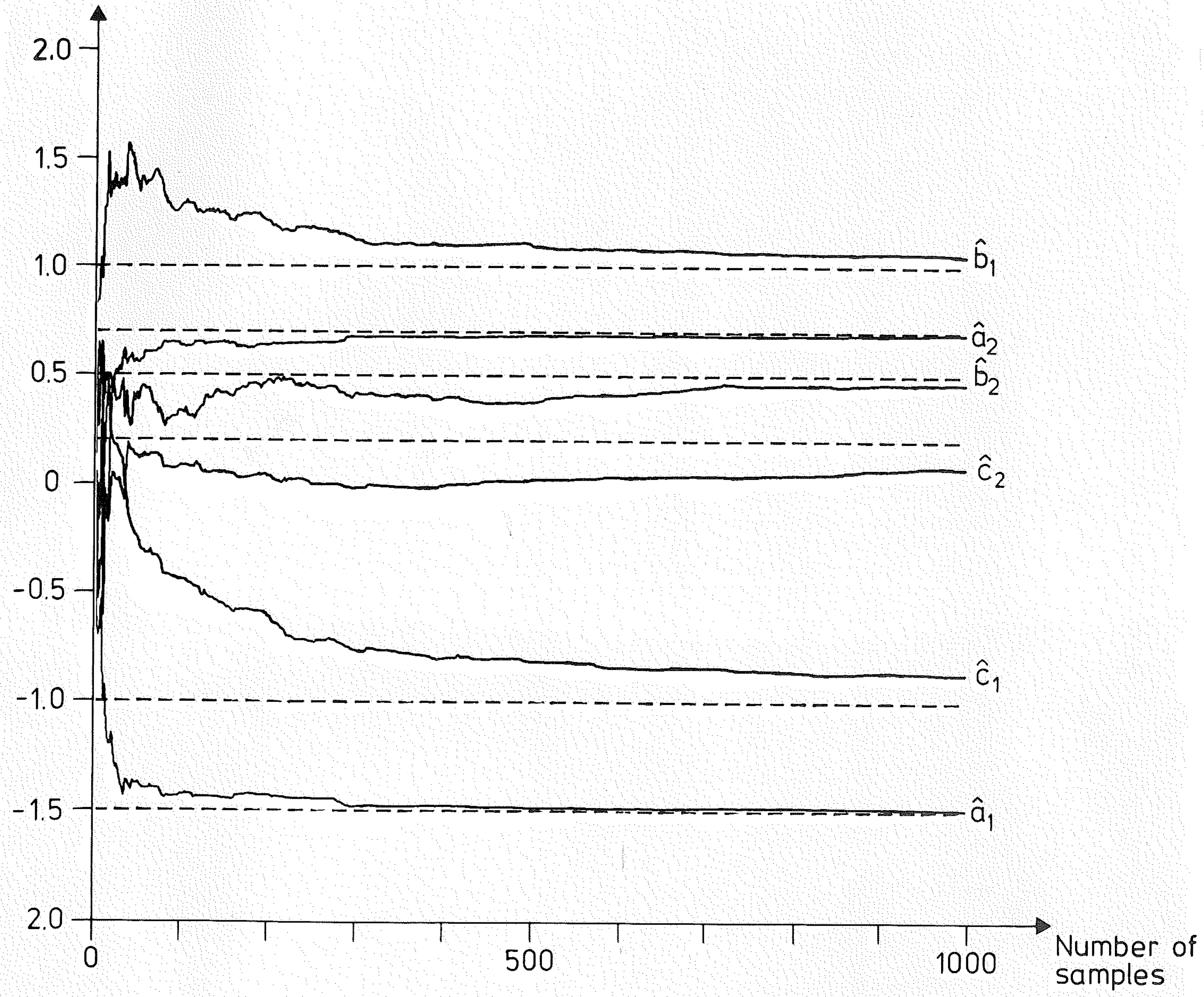


Figure 5.1 Parameter estimates of a second order system. Straightforward application of the algorithm is done. The dashed lines give the true values of the parameters.

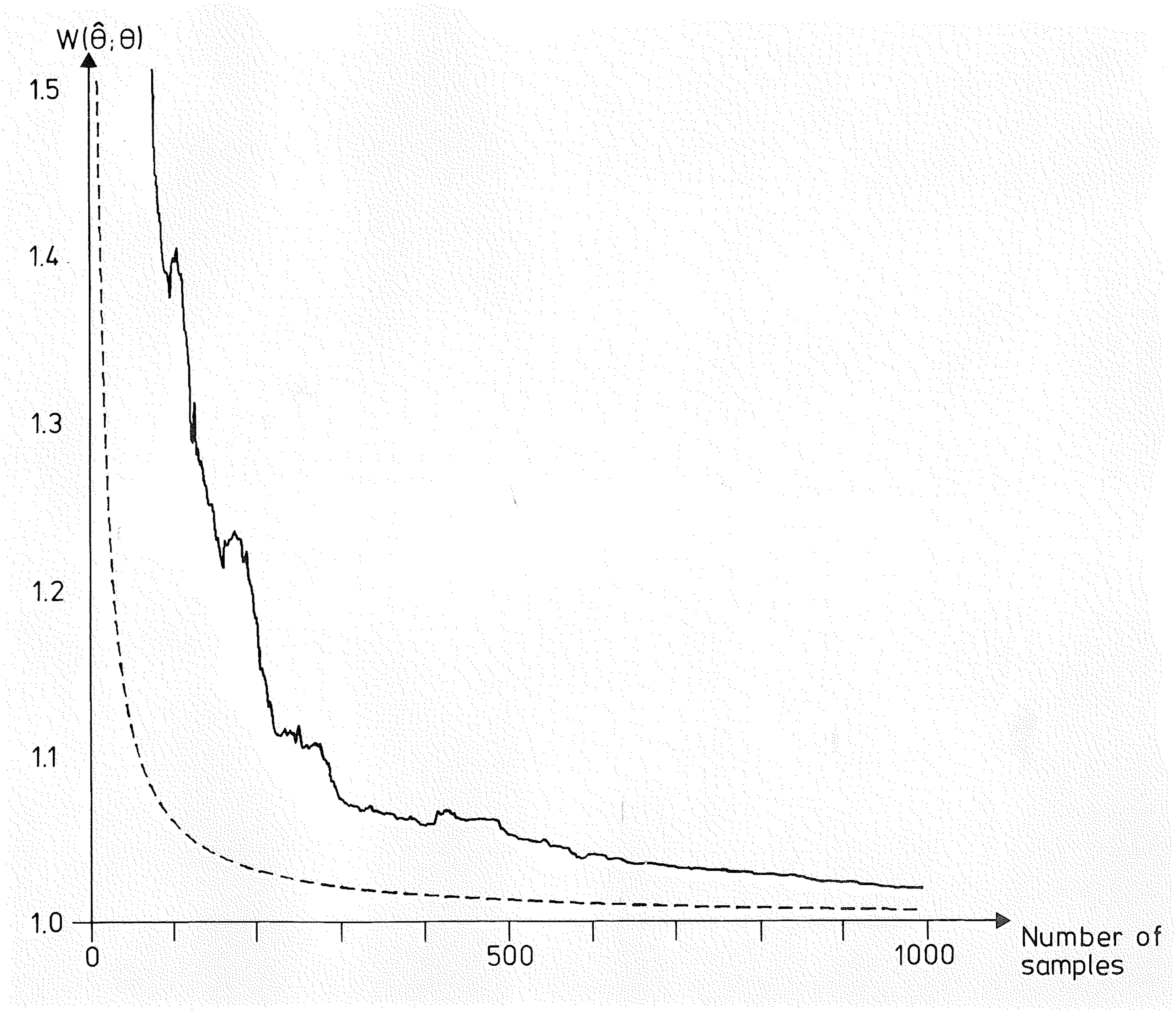


Figure 5.2 Loss function of a second order system. Straight-forward application of the algorithm is done. The dashed line gives the asymptotically expected loss.

will be considerably better. If this idea would have any practical value $\hat{\theta}$ must not change too much in the rest of the identification.

The second idea has been examined to some extent by simulations. The particular system given in table 2 was used. In some simulations $\hat{\theta}_t$ was fixed to the correct values for the first 100 samples and after that $\hat{\theta}$ was estimated according to the algorithm. Good results were obtained. In other simulations the first 100 samples were used at an identification with the off-line ML algorithm. These off-line identifications produced good initial values of the recursive algorithm, which produced satisfactory results in this case.

The algorithm has been modified in the following way. It is applied straight-forward in N_1 steps. Then a test of "convergence" is performed. If "convergence" has occurred; the algorithm is continued straight-forward. If no convergence has occurred a restart is made with $\hat{\theta}_t$ keeping its value and the other variables as their ordinary start values. $\hat{\theta}_t$ is constrained to be constant for the next N_2 steps. After another N_1 steps a new test of "convergence" is made. The estimate $\hat{\lambda}$ is modified in an obvious way with regard to the latest restart. This prodedure of successive restarts is continued until "convergence" has occurred.

A suitable test of "convergence" would be to use $W(\hat{\theta};\theta)$. If this quantity is small (close to 1) "convergence" may be considered to have occurred. However, this test quantity cannot be used in practice, since it requires knowledge of θ and λ^2. Instead $W(\hat{\theta}_t;\hat{\theta}_{t-N_1-N_2})$ is used with λ^2 substituted by $\hat{\lambda}_t^2$. This means that θ is substituted by the estimate $\hat{\theta}$ which was present when the latest test of "convergence" was made. If the test quantity is smaller than VTEST no more restarts are made.

Simulations were made using the same realizations as before. The values of the variables used were VTEST 1.05, $N_1 = 300$, and $N_2 = 50$. It is the author's experience that the method is not very sensitive to the values of the parameters VTEST, N_1 and N_2. The results are good for RMLE1 and a bit inferior, but yet satisfactory for RMLE2, see table 5.3.

	a_1	a_2	b_1	b_2	c_1	c_2	λ	W
Expected mean	-1.5	0.7	1.0	0.5	-1.0	0.2	1.0	1.0030
RMS error	0.007	0.006	0.022	0.029	0.023	0.022	0.032	0.0032
RMLE1 mean	-1.498	0.699	0.998	0.500	-0.987	0.180	0.994	1.0057
RMS error	0.008	0.008	0.022	0.025	0.034	0.042	0.048	0.0065
RMLE2 mean	-1.505	0.702	1.002	0.479	-0.966	0.160	1.009	1.017
RMS error	0.014	0.014	0.032	0.062	0.062	0.080	0.060	0.020

Table 5.3. Results for a second order system. The trick with restarts is used. RMLE1 is the general algorithm given in chapter II. RMLE2 is Young's algorithm.

In Figures 5.3 and 5.4 it is shown how the modified algorithm RMLE1 works on the same data as were used in Figures 5.1 and 5.2.

It can be seen that the restarts give the algorithm larger "gain" than before which causes a jerkiness of the estimates. The long range effect, however, is that the estimates are considerably closer to the correct values than without restarts.

Now a brief discussion of other tricks and approaches tried by the author is given. His experience is that these tricks do not give a satisfactory improvement of the algorithm.

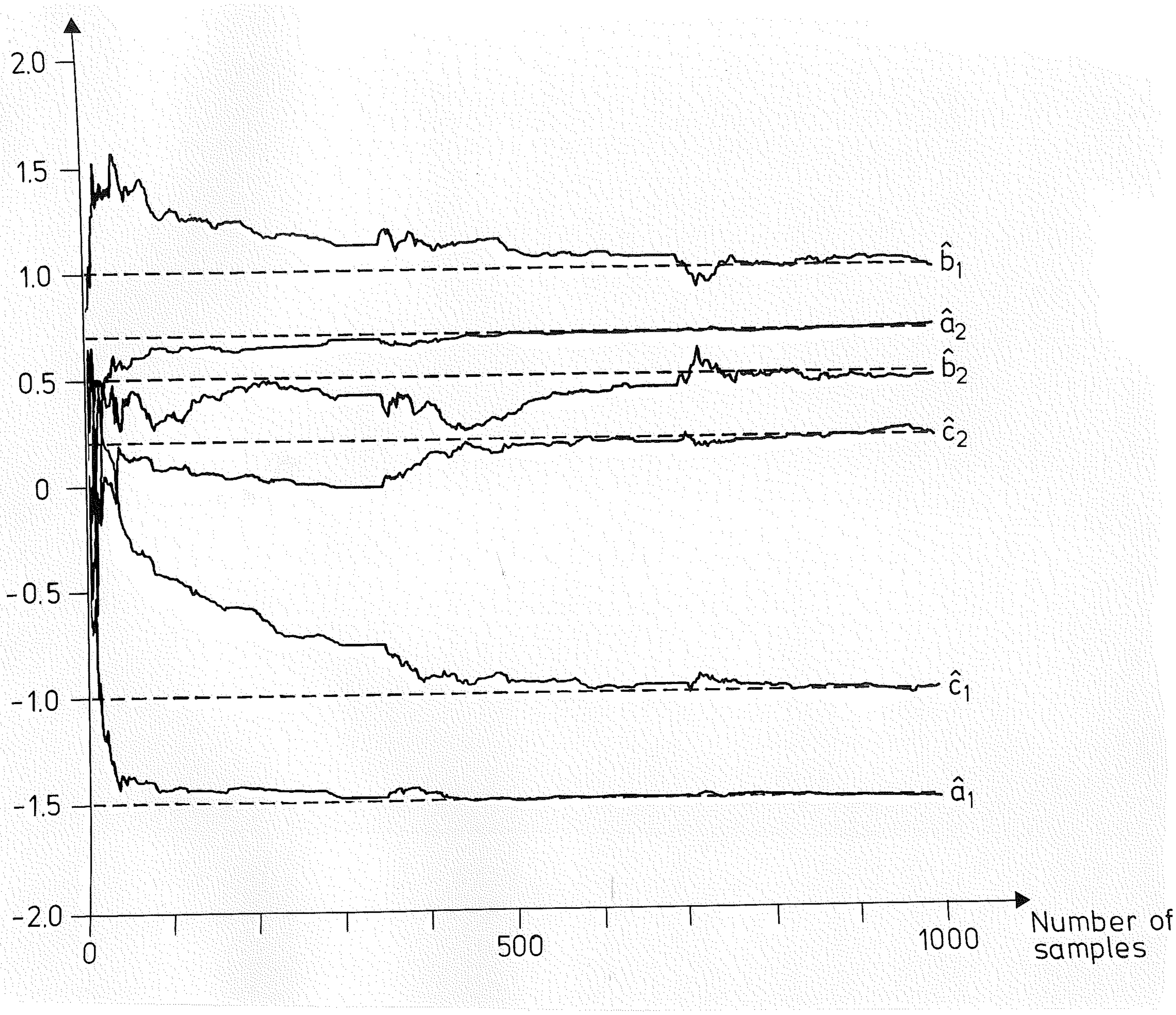


Figure 5.3 Parameter estimates of a second order system. The modified algorithm is used. The dashed lines give the true values of the parameters.

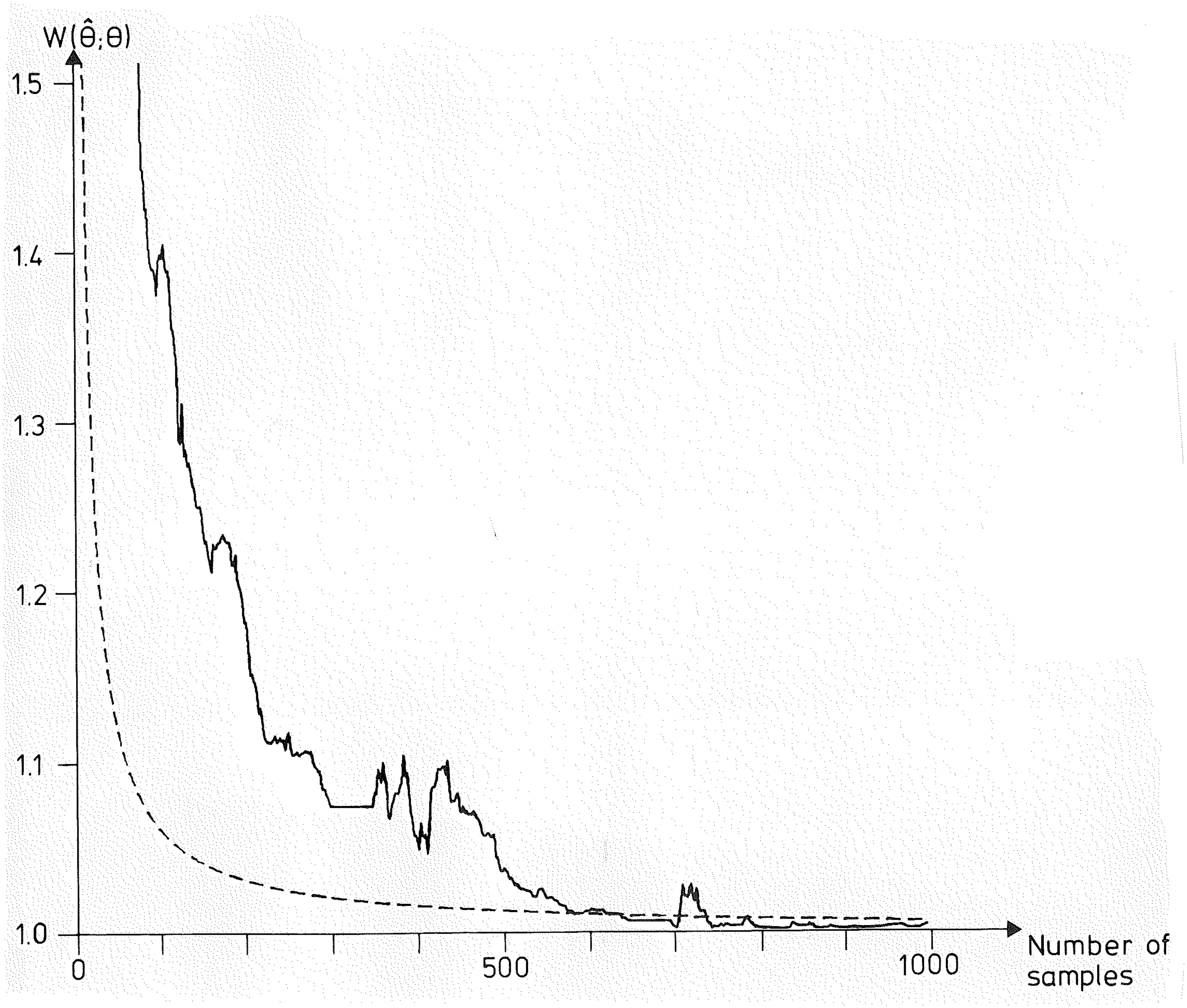


Figure 5.4 Loss function of a second order system. The modified algorithm is applied. The dashed line gives the asymptotically expected loss.

- The inverse $V_N''(\hat{\theta}_{N-1})^{-1}$ was computed by inversion of $V_N''(\hat{\theta}_{N-1})$, i.e. (2.12) was used together with inversion instead of (2.18). Since $V_N''(\hat{\theta})$ is not independent of $\hat{\theta}$ this may change the result of the algorithm.

- The term $\varepsilon(N+1;\hat{\theta}_N)\varepsilon''(N+1;\hat{\theta}_N)$ was not dropped in the computation of $V_{N+1}''(\hat{\theta}_N)$.

- If the algorithm does not really minimize V_N the approximation $V_N'(\hat{\theta}_N) = 0$ may not be accurate. The equation (2.11) of the gradient was changed to

  $$V_{N+1}'(\hat{\theta}_N) = \alpha V_N'(\hat{\theta}_N) + \varepsilon(N+1;\hat{\theta}_N)\varepsilon'(N+1;\hat{\theta}_N)$$

  The parameter α was chosen in the interval [0, 1]. When $\alpha = 0$ the previous algorithm is obtained. The choice of $\alpha = 1$ caused very large changes in the parameter estimates and was very unsatisfactory. The choice $\alpha = 0.6$ gave some improvements of the convergence but it was not satisfactory enough.

- In order to speed up the convergence it may be appropriate to change (2.17) to

  $$\hat{\theta}_{N+1} = \hat{\theta}_N - N^{\beta}P_{N+1}\varphi_{N+1}\varepsilon_{N+1}$$

  where $0 \leq \beta < 1$. This attempt gave no improvement in a few simulated examples.

- The normalized loss function $\frac{1}{N}V_N(\hat{\theta})$ was minimized instead of $V_N(\hat{\theta})$. No significantly improvements occured.

VI. APPLICATION TO PLANT MEASUREMENTS

It was shown in chapter V that the recursive ML method worked well on the simulated data. In order to examine the properties of the algorithm when it is applied to real data, plant measurments were used. Measurements from a nuclear reactor and from a laboratory heat diffusion process were tried. Identification using an off-line ML algorithm on the same data have been made by others. Comparisons are made between the results of the different methods.

It turned out that it is much more difficult to get the algorithm to work satisfactorily on real data. There are probably several reasons for that, for example that the structure and the order of the process is not known.

Different values of the parameters N_1, N_2, and VTEST were tried. The results of the identifications were not very sensitive to the choice of these values. However, it cannot be excluded that better results may be possible to obtain by other choices or by a suitable combination of the tricks mentioned in chapter V.

To illustrate the on-line identification procedure the estimates $\hat{\theta}_t$ and the residuals $\hat{\varepsilon}_t$ are plotted versus the time t. The comparison between the results of the off-line and the on-line algorithms are illustrated with plots of the following signals:

1. the input $u(t)$
2. the output $y(t)$
3. the model output $y_m(t) = \dfrac{\hat{B}(q^{-1})}{\hat{A}(q^{-1})}\, u(t)$
4. the model error $e_m(t) = y(t) - y_m(t)$
5. the residuals $\varepsilon(t;\hat{\theta})$

The model outputs of the on-line models were computed using the parameter estimates obtained at the last sampling interval in the identification.

Example 1

The system is a nuclear power reactor in Ågesta, Sweden. The data were supplied to the Division of Automatic Control by AB Atomenergi, Studsvik, Sweden. The system is described briefly in Gustavsson (1969a) where also ML identifications are reported. The number of data is 1700 and the measurements are called AR 60. The input is control rod position and the output is the nuclear power. An idealized input signal was used both for on-line and off-line identification. The sampling interval is 1 second. Using an F-test it is concluded in Gustavsson (1969a) that the system is of third order.

When recursive ML identification was performed for a model of third order several problems arose. The parameter estimates did not converge. At no time their values were close to the parameter values obtained in Gustavsson (1969a). However, the model outputs of the two models did not differ significantly. A possible explanation of these phenomena is that the order of the model was chosen too high. An indication of this is that both the model in Gustavsson (1969a) and the model obtained by on-line identification have approximately one pole and zero in common.

The results of the on-line identification of a second order model were more satisfactory. The parameters N_1, N_2, and VTEST were chosen as 300, 50 and 1.05 respectively. The parameter estimates obtained are given in Table 6.1. In Gustavsson (1969a) 95 % confidence intervals of the parameter estimates are given. Only the parameters $\hat{a}_1$ and $\hat{c}_2$ of the model obtained on-line are inside these intervals.

	On-line algorithm used	Off-line algorithm used
$\hat{a}_1$	-0.95	-1.08
$\hat{a}_2$	0.14	0.20
$\hat{b}_1$	1.69	1.69
$\hat{b}_2$	-1.12	-1.31
$\hat{c}_1$	-0.76	-0.92
$\hat{c}_2$	0.23	0.27
$\hat{\lambda}$	0.18	0.17

Table 6.1 Results of identification of the nuclear reactor data.

Figure 6.1 shows how the parameter estimates $\hat{\theta}_t$ and the estimated residuals $\hat{\varepsilon}_t$ vary with time. The large values of $\hat{\varepsilon}_t$ at t = 300, 650, 1000 and 1350 are due to the restarts. The large residuals at t = 41, 143, 1233, 1291, 1517, and 1597 are explained by large measurement errors at these points. The measurement errors can be seen clearly from plots of the data.

In Figures 6.2 the model outputs and the residuals are shown for different models. When the second order models are compared it is clear from Table 6.1, Figures 6.1 and 6.2 that there are only small differences between the results of on-line identification and the result of off-line identification. The model output for a third order model computed by on-line identification was very similar to the

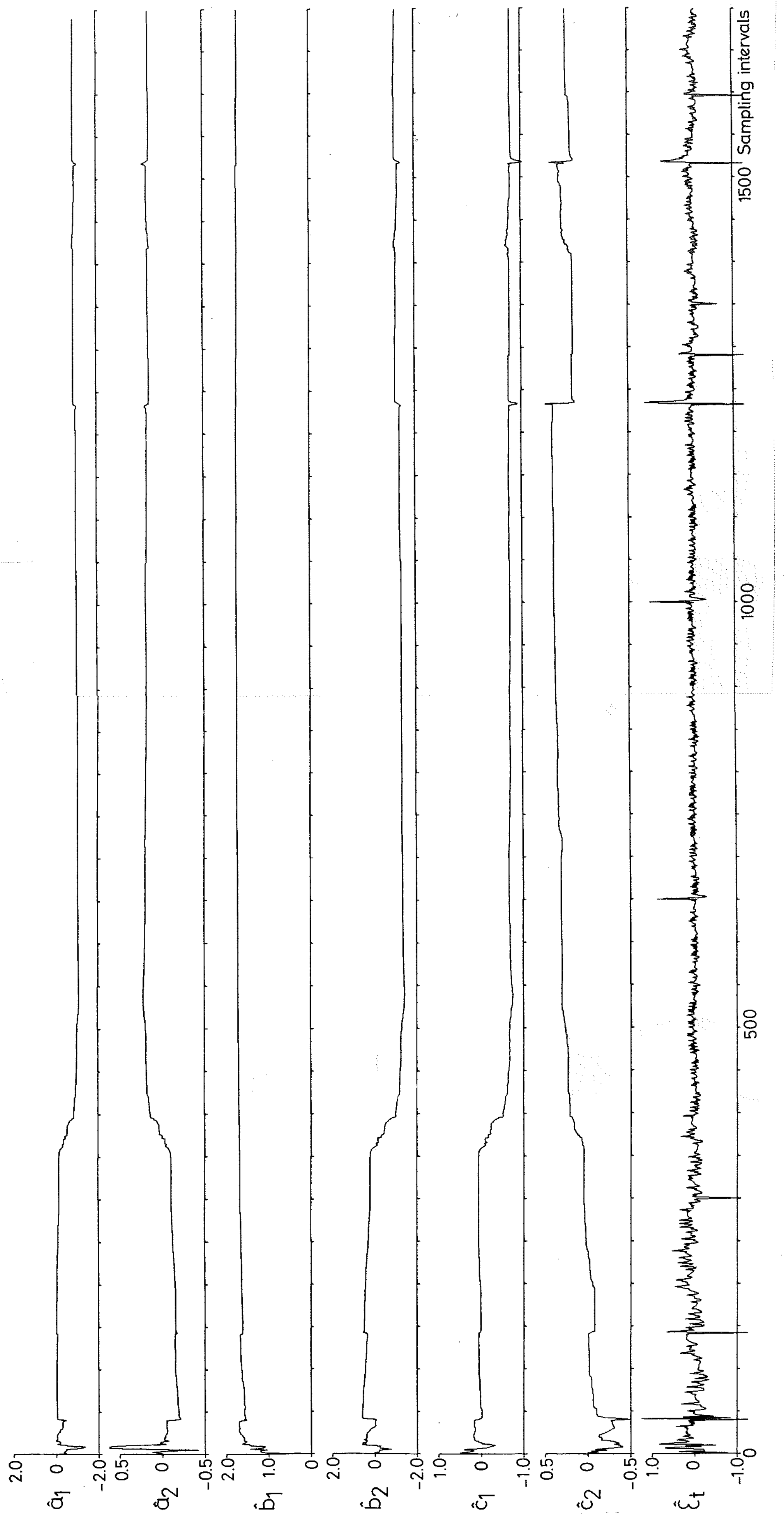


Figure 6.1 The parameter estimates and the residuals estimated for the nuclear reactor data. The sampling interval is 1 second.

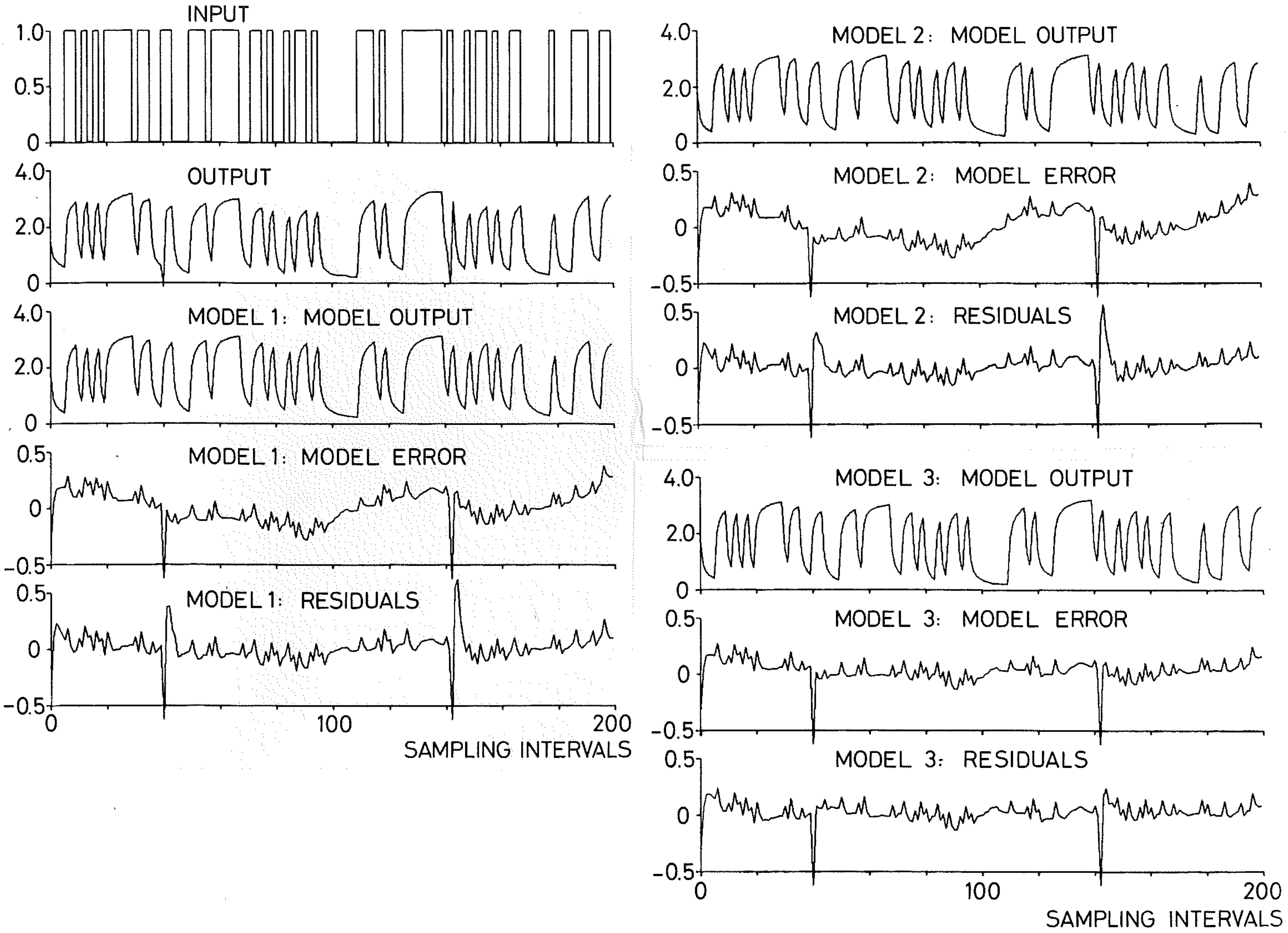


Figure 6.2 Three models of the nuclear reactor.
Model 1 is of second order and is obtained by on-line identification.
Model 2 is of second order and is obtained by off-line identification.
Model 3 is of third order and is obtained by off-line identification.
Digital units are used. The sampling interval is 1 second.
Notice the different scales.

model outputs of the second order models. The best result is obtained with a third order model obtained by off-line identification. However, the improvements are not very great as can be seen in Figure 6.2. The slow oscillation with small amplitude in the model error disappears, however.

Example 2

The system is a laboratory heat diffusion process at the Division of Automatic Control, Lund Institute of Technology. The process consists of a long copper rod. The end temperatures can be controlled using Peltier elements. Identification results of the system using the off-line ML method as well as a short description of the process is given in Leden (1971). The data used here are called series S1. The input is the temperature of one of the end points of the rod. The other end point temperature was kept constant. The output of the process is the temperature in the middle of the rod. The number of data is 862 and the sampling interval is 10 seconds. Leden (1971) found that a model of fourth order was appropriate.

Recursive identification was performed with $N_1 = 200$, $N_2 = 50$,and VTEST = 1.05. The resulting parameter estimates are given in Table 6.2. They differ very much from the estimates obtained with off-line identification. In Figure 6.3 it is shown how the estimates vary with time. The large values of the residuals at t = 200 and 450 are due to the restarts.

The model identified off-line is obtained by a straight-forward application of the ML algorithm. In Leden (1971) also a considerably better model is obtained by inclusion of estimation of initial values and constant errors and by limiting the residuals. This improved model has four real-valued poles and the model error is much smaller than before.

	On-line algorithm used	Off-line algorithm used
$\hat{a}_1$	-0.88	-2.03
$\hat{a}_2$	-0.35	1.40
$\hat{a}_3$	-0.01	-0.40
$\hat{a}_4$	0.27	0.04
$\hat{b}_1 \cdot 10^3$	1.08	0.02
$\hat{b}_2 \cdot 10^3$	1.22	0.46
$\hat{b}_3 \cdot 10^3$	4.43	3.90
$\hat{b}_4 \cdot 10^3$	8.87	2.30
$\hat{c}_1$	0.44	-0.86
$\hat{c}_2$	0.32	0.54
$\hat{c}_3$	0.26	-0.15
$\hat{c}_4$	-0.03	0.24
$\hat{\lambda} \cdot 10^3$	2.53	0.36

Table 6.2 Results of identification of the heat rod data.

In Figure 6.4 the model identified on-line and the model identified by a straight-forward off-line ML algorithm are compared. These two models differ very much in the parameter values. It can be seen from Figure 6.4, however, that the model obtained by on-line identification describes the slowest modes of the process well. When the input is constant for a longer period the residuals are small. The fast modes of the process are badly estimated.

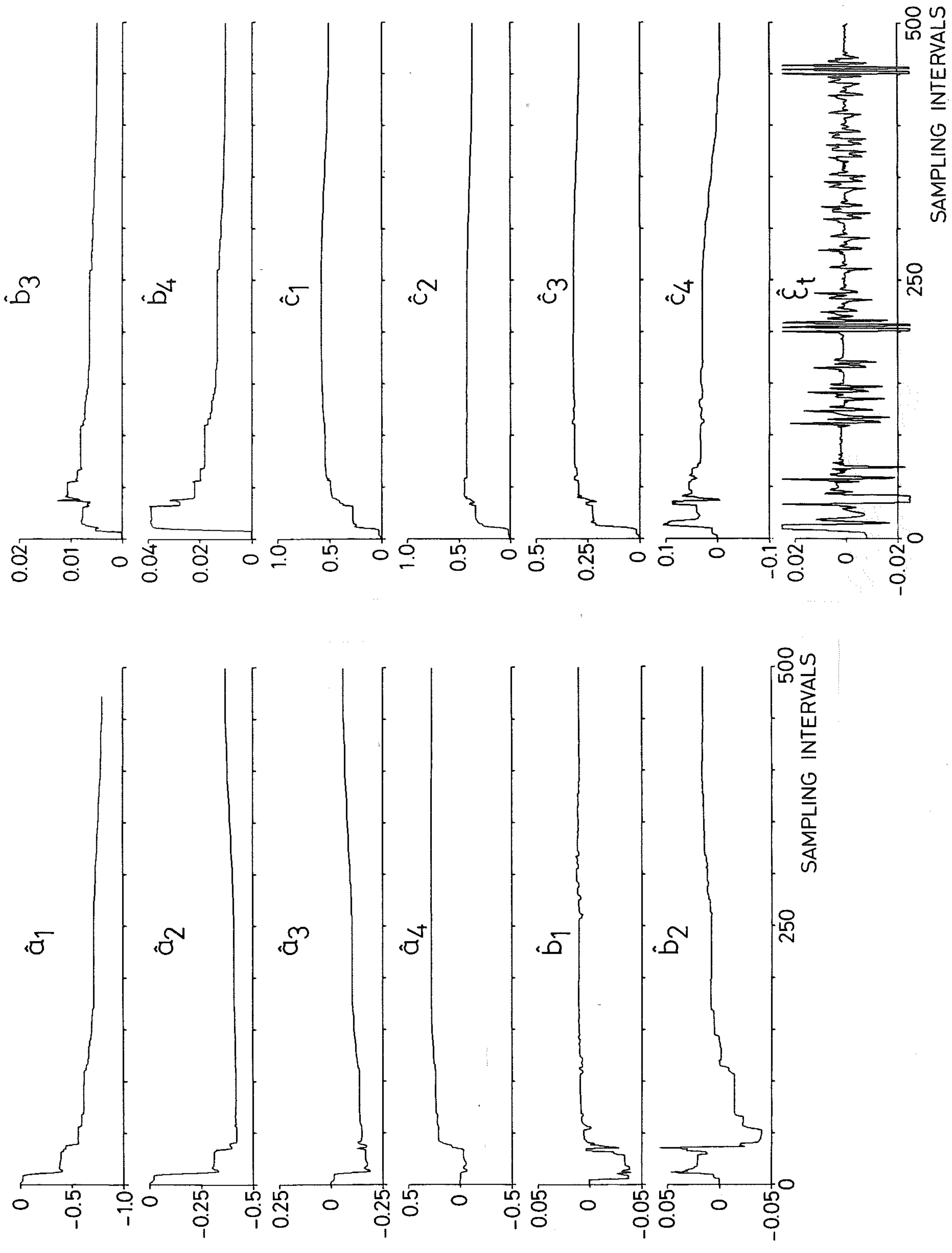


Figure 6.3 The parameter estimates and the residuals estimated for the heat rod data. The sampling interval is 10 seconds.

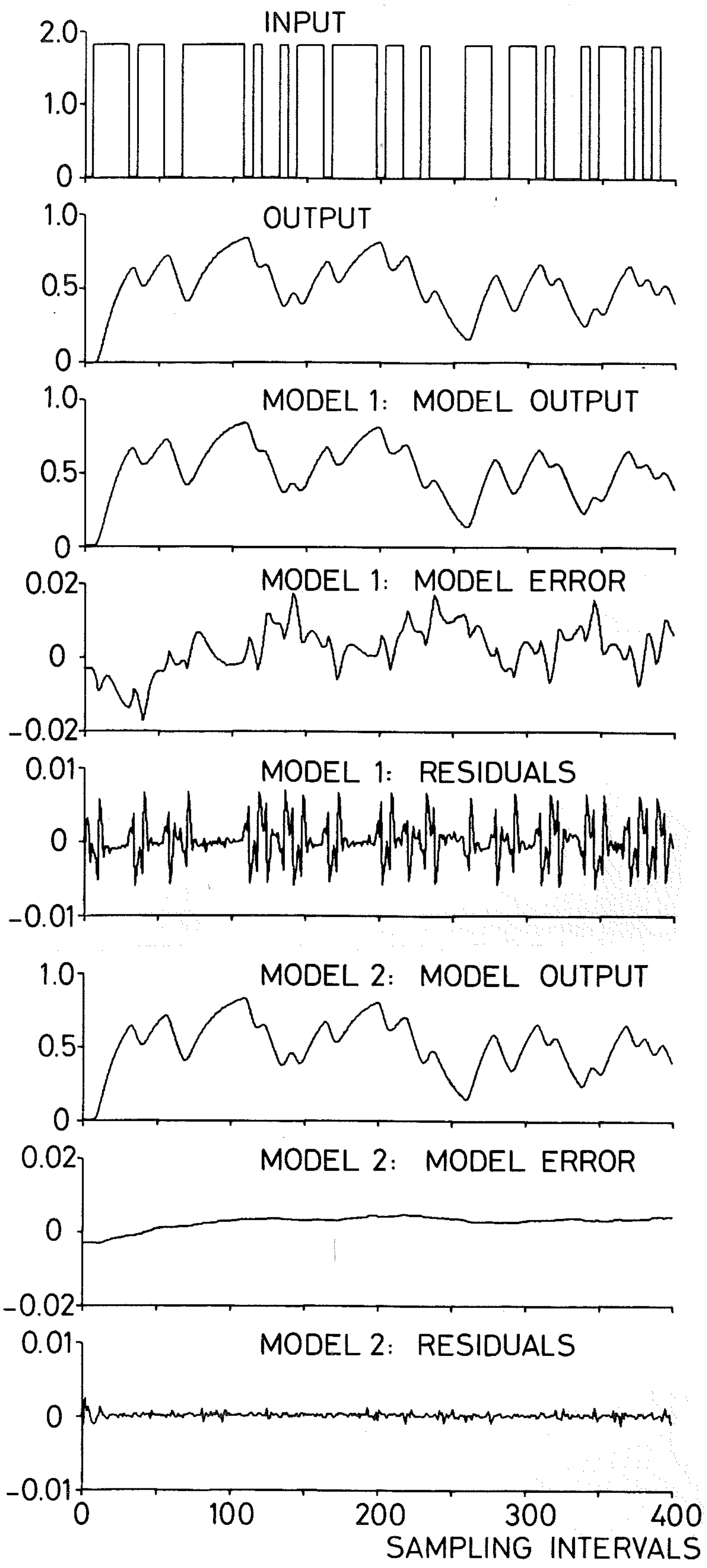


Figure 6.4 Models of the heat diffusion process.

Model 1 is obtained by on-line identification.
Model 2 is obtained by off-line identification.

All variables are given in °C. Constant levels are added to the input, the output and the model outputs. The sampling interval is 10 seconds. Notice the different scales.

CONCLUSIONS

The following conclusions are based on the examples given.

- The algorithm must be applied with caution. It does not give as good estimates as the off-line ML algorithm.

- Applied to simulated data the algorithm works quite well when suitable tricks are used.

- Applied to real data it is difficult to get the algorithm working satisfactorily. An improper choice of the order of the model may cause considerable difficulties. The most dominating modes of the process are well estimated. It is probably often appropriate to use a low order model.

- The choice of the values of N_1, N_2, and VTEST is not very crucial.

ACKNOWLEDGEMENTS

The author wants to express his great gratitude to Professor K J Åström. He has done a first examination of the algorithm. His guidance was very valuable. Tekn lic Ivar Gustavsson has continued professor Åström's preliminary work. I owe him my sincere thanks for a good cooperation and enlightening discussions.

It is also a pleasure to thank Miss Kerstin Palmqvist who typed the manuscript and Mr Bengt Lander who prepared the figures.

The author is grateful for the data which were supplied to the Division of Automatic Control by AB Atomenergi and Tekn lic Bo Leden, Lund.

This project was partially supported by the Swedish Board of Technical Development under contract 72-202/U137.

REFERENCES

Åström, K.J. (1968).
Lectures on the Identification Problem - the Least Squares Method. Report 6806, Division of Automatic Control, Lund Institute of Technology.

Åström, K.J. - Bohlin, T. (1966).
Numerical Identification of Linear Dynamic Systems from Normal Operating Records. Paper, IFAC Symposium on Theory of Self-Adaptive Systems, Teddington, England. Also in Theory of Self-Adaptive Control Systems (Ed. P.H. Hammond), Plenum Press, New York.

Åström, K.J. - Eykhoff, P. (1971).
System Identification - A Survey. Automatica 7, 123 - 162.

Åström, K.J. - Söderström, T. (1973).
Uniqueness of the Maximum Likelihood Estimates of the Parameters of a Mixed Autoregressive Moving Average Process, Report 7306, Division of Automatic Control, Lund Institute of Technology.

Bohlin, T. (1970).
On the Maximum Likelihood Method of Identification. IBM J. Res. and Dev., 14, No 1, 41 - 51.

Clarke, D.W. (1967).
Generalized Least Squares Estimation of the Parameters of a Dynamic Model. 1st IFAC Symposium on Identification in Automatic Control Systems. Prague.

Gustavsson, I. (1969a).
Maximum Likelihood Identification of Dynamics of the Ågesta Reactor and Comparison with Results of Spectral Analysis. Report 6903, Division of Automatic Control, Lund Institute of Technology.

Gustavsson, I. (1969b).
Parametric Identification on Multiple Input, Single Output Linear Dynamic Systems. Report 6907, Division of Automatic Control, Lund Institute of Technology.

Leden, B. (1971).
Identification of Dynamics of a One Dimensional Heat Diffusion Process. Report 7121, Division of Automatic Control, Lund Institute of Technology.

Ljung, L. (1973).
New Convergence Criteria for Stochastic Approximation Algorithms. Forthcoming report. Division of Automatic Control, Lund Institute of Technology.

Ljung, L. - Wittenmark, B. (1973).
Asymptotic Properties of Self-Tuning Regulators Based on Least Squares Identification. Forthcoming report. Division of Automatic Control, Lund Institute of Technology.

Panuška, V. (1968).
A Stochastic Approximation Method for Identification of Linear Systems Using Adaptive Filtering. 1968 JACC, Ann Arbor, Michigan.

Söderström, T. (1972).
On the Convergence Properties of the Generalized Least Squares Identification Method. Report 7228, Division of Automatic Control, Lund Institute of Technology.

Söderström, T. (1973).
On the Uniqueness of Maximum Likelihood Identification for Different Structures. Report 7307, Division of Automatic Control, Lund Institute of Technology.

Valis, J. - Gustavsson, I. (1969).
Some Computational Results Obtained by Panuška's Method of Stochastic Approximations for Identification of Discrete Time Systems. Report 6915, Division of Automatic Control, Lund Institute of Technology.

Wieslander, J. (1971).
Real Time Identification, Part I. Report 7111, Division of Automatic Control, Lund Institute of Technology.

Young, P.C. (1970).
An Extension to the Instrumental Variable Method for Identification of a Noisy Dynamic Process. Univ. of Cambridge, Dep of Eng, Technical note CN/70/1.

Young, P.C. - Shellswell, S.H. - Neethling, C.G. (1971).
A Recursive Approach to Time Series Analysis. Univ. of Cambridge, Dep of Eng, CUED/B - Control/TR16.

APPENDIX

The purpose of this appendix is to show that the equation (4.2) can be substituted by (4.3). The basic tool is the following lemma which is taken from Ljung (1973).

Lemma. Let $\{f_n\}$ be a strictly stationary process such that $E|f_n|$ exists.Assume that the sequence $\{a_n\}$ fulfils

$$a_n \to 0 \quad \text{a.s. } n \to \infty$$

Then

$$\frac{1}{N}\sum_{i=1}^{N} a_i f_i \to 0 \quad \text{a.s. } N \to \infty$$

Corr 1. Let $\{\theta_n\}$ be a sequence of stochastic variables such that

$$\theta_n \to \theta^* \quad \text{a.s. } n \to \infty$$

Further let $\{f_n(\theta)\}$ be a strictly stationary process, which depends on the parameter θ such that $f_n(\theta)$ is (continuously) differentiable with respect to θ a.s. and that $E|f_n'(\theta)|^2$ exists if θ belongs to some neighbourhood of θ^*.

Assume that the sequence $\{a_n\}$ is bounded a.s. Then

$$\frac{1}{N}\sum_{i=1}^{N} [f_i(\theta_i) - f_i(\theta^*)]a_i \to 0 \quad \text{a.s.} \quad N \to \infty$$

Proof The assumptions imply

$$|f_i(\theta_i) - f_i(\theta^*)| \leq M_i|\theta_i - \theta^*| \qquad i \geq N_0$$

for some N_0, where $\{M_i\}$ is a strictly stationary process such that $E|M_i|$ exists. Thus

$$\left|\frac{1}{N}\sum_{i=1}^{N}[f_i(\theta_i) - f_i(\theta^*)]a_i\right| \leq$$

$$\leq \frac{1}{N}\sum_{i=1}^{N_0}|f_i(\theta_i) - f_i(\theta^*)|\ |a_i| + \frac{1}{N}\sum_{i=N_0+1}^{N}|f_i(\theta_i) - f_i(\theta^*)||a_i|$$

$$\leq \frac{1}{N}\sum_{i=1}^{N_0}|f_i(\theta_i) - f_i(\theta^*)||a_i| + \frac{1}{N}\sum_{i=1}^{N}M_i|\theta_i - \theta^*||a_i|$$

The first term trivially tends to zero as N tends to infinity. It follows from the lemma that the second term tends to zero as well.

Corr 2. Let the assumption of $\{a_n\}$ in Corr 1 be changed. Assume instead that $\{a_n\}$ is a strictly stationary process such that $E|a_n|^2$ exists. Then the result of Corr 1 remains true.

In the present algorithm ε_t and φ_t are computed in an approximate way as discussed in chapter II. To simplify the calculations it will be assumed here that they are computed exactly. The results of Ljung-Wittenmark (1973) indicate that it may be possible to extend the calculations to the actual ε_t and φ_t.

Assumptions on the distribution of the noise will be made indirectly. It will be assumed that the expectations

$$E|\varepsilon'(t;\hat{\theta})|^2 \qquad \text{and} \qquad E|\varepsilon''(t;\hat{\theta})|^2$$

exist for all $\hat{\theta}$ such that the corresponding polynomials

$\hat{A}(z)$ and $\hat{C}(z)$ have all zeros outside the unit circle.

The residuals $\varepsilon(t;\hat{\theta})$ and the gradient $\varepsilon'(t;\hat{\theta})$ are strictly stationary processes if the initial values are chosen properly. However, the effect of the initial values do not affect the result and it will be assumed generally that they are chosen in a proper way.

To simplify,the following notations will be used

$$\varphi_t = \varepsilon'(t;\hat{\theta}_{t-1}) \qquad \varphi_t^* = \varepsilon'(t;\theta^*)$$

$$\varepsilon_t = \varepsilon(t;\hat{\theta}_{t-1}) \qquad \varepsilon_t^* = \varepsilon(t;\theta^*)$$

The calculations are organized as proofs of three assertions.

<u>Assertion 1</u> $\lim\limits_{N\to\infty} \frac{1}{N}\sum\limits_{t=1}^{N} \varphi_t\varphi_t^T\hat{\theta}_N = (\lim\limits_{N\to\infty}\frac{1}{N}\sum\limits_{t=1}^{N} \varphi_t\varphi_t^{*T})\theta^*$ a.s.

<u>Proof</u> After a decomposition the sum of the left hand side is written as

$$\frac{1}{N}\sum_{t=1}^{N} \varphi_t\varphi_t^T\hat{\theta}_N = \frac{1}{N}\sum_{t=1}^{N} (\varphi_t^*\varphi_t^{*T})\theta^*$$

$$+ \frac{1}{N}\sum_{t=1}^{N} (\varphi_t\varphi_t^T-\varphi_t^*\varphi_t^{*T})\hat{\theta}_N + \frac{1}{N}\sum_{t=1}^{N} \varphi_t^*\varphi_t^{*T}(\hat{\theta}_N-\theta^*)$$

It follows from Corr 1 that the second term tends to zero and from the lemma that the third term tends to zero.

□

Assertion 2 $\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\varphi_t\varphi_t^T\hat{\theta}_{t-1} = (\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\varphi_t^*\varphi_t^{*T})\theta^*$ a.s.

Proof A decomposition gives

$$\frac{1}{N}\sum_{t=1}^{N}\varphi_t\varphi_t^T\hat{\theta}_{t-1} = \frac{1}{N}\sum_{t=1}^{N}(\varphi_t^*\varphi_t^{*T})\theta^*$$

$$+\frac{1}{N}\sum_{t=1}^{N}(\varphi_t\varphi_t^T - \varphi_t^*\varphi_t^{*T})\hat{\theta}_{t-1} + \frac{1}{N}\sum_{t=1}^{N}\varphi_t^*\varphi_t^{*T}(\hat{\theta}_{t-1}-\theta^*)$$

Using the same type of arguments as in the preceding proof the assertion follows.

□

Assertion 3 $\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\varepsilon_t\varphi_t = \lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\varepsilon_t^*\varphi_t^*$ a.s.

Proof Using a similar decomposition as before

$$\frac{1}{N}\sum_{t=1}^{N}\varepsilon_t\varphi_t = \frac{1}{N}\sum_{t=1}^{N}\varepsilon_t^*\varphi_t^* + \frac{1}{N}\sum_{t=1}^{N}\varepsilon_t^*(\varphi_t-\varphi_t^*)$$

$$+\frac{1}{N}\sum_{t=1}^{N}(\varepsilon_t-\varepsilon_t^*)\varphi_t^* + \frac{1}{N}\sum_{t=1}^{N}(\varepsilon_t-\varepsilon_t^*)(\varphi_t-\varphi_t^*)$$

The second and the third terms tend to zero according to Corr 2. It follows from the lemma (put f_n=1) that the fourth term tends to zero.

□

It can be shown, see e.g. Söderström (1972), that the right hand sides of the assertions really exist under mild conditions.

CORRECTIONS

The abbreviation pa.b denotes page a line b.

p6.10 Read "$n_a < i \leq n_b$"

p9.6 Read "the coefficients b_{n_b}, a_{n_a}"

p11.4 Read "given by dim R(P)"

p20.4,p20.5 Read "three times differentiable"

p20.6 Read "fixed parameter"

p22.15, p24.1, p24.4, p25.3. p29.8 Read "= 0 a.s."

p24.10 Read "$\hat{A}(q^{-1}) \equiv \hat{A}(q^{-1})$"

p24.13 Read "since trivial calculations show that V is strictly convex"

p26.3, p27.13, p28.16, p31.9, p32.18, p40.14, p42.1, p46.18, p46.19, p47.11, p48.10, p52.14. The equality $\hat{\theta} = \theta$ is not consistent if the vectors are of different orders. The meaning is for p26.3 $\hat{A}(q^{-1}) \equiv A(q^{-1})$, $\hat{B}(q^{-1}) \equiv B(q^{-1})$, $\hat{C}(q^{-1}) \equiv C(q^{-1})$
For the other cases the modifications are analogous.

p27.6 Replace "=" with "≡"

p28.14, p30.8 Read "Lemma 2.4 Corr and Lemma 2.6"

p30.16 Replace the line with "which gives"

p32.16 Read "Lemma 2.1, Lemma 2.4 Corr and Lemma 2.6"

p33.3 Read "measurement"

p37.19 Read "$\overline{B}(q^{-1})=$"

p45.16 Read "$\hat{A}(q^{-1}) \equiv \tilde{A}(q^{-1})\hat{L}(q^{-1})$, $\hat{B}(q^{-1}) \equiv \tilde{B}(q^{-1})\hat{L}(q^{-1})$"

p46.8 Read "a unique stationary point"

p47.6 Add "and $\hat{n}_b \geq n_b$"

p52.5 Read "minimum point with respect to $(\hat{a}_1 \ldots \hat{a}_{\hat{n}_a}\ \hat{b}_1 \ldots \hat{b}_{\hat{n}_b})$"

p55.9 Read "Canonical"

pB.3.11 Read "Kaufman"

pB.8.13 Read "number"

CORRECTIONS

The abbreviation pa.b means page a line b.

p3.3	Read "the integration path"
p10.7	Replace "$D^{(t_i - 1 - \nu)}$" with "$D^{(t_k - 1 - \nu)}$"
p10.12 and p13.2	Replace "$\prod_{j \neq k} (z - u_j)^{t_k}$" with "$\prod_{j \neq k} (z - u_j)^{t_j}$"
p13.12	Read "$\prod_{k=1}^{\ell} (z - u_k)^{t_k}$"
p15.13	Read "$\hat{C}'(z) \equiv \hat{C}(z)$"
p15.20	Delete "that"

CORRECTIONS

The abbreviation pa.b denotes page a, line b.

p1.2 Delete "the"

p7.5 Read "derived for the LS case"

p8.17 Read "residuals"

p16.Table5.1 The theoretical RMS error of λ is 0.016

p17.Table5.2 and p21.Table5.3 The theoretical RMS error of λ is 0.016 and the theoretical RMS error of W is 0.035

p17.7 and p20.5 Read "table 5.2"

p19 The scale on the W-axis is incomplete. Figure 5.4 shows the correct scale.

p25.14 Read "are not known"

p26.31 Replace "$\hat{a}_1$" with "$\hat{b}_1$"

p41.11 Replace "φ_t" in the right hand side with "φ_t^*"